Who Benefits from Reduced Social Media Use? A Causal Analysis of Heterogeneous Treatment Effects on Psychological Well-being

Lily Sarfati Robinson 300680676
02/06/2025
Victoria University of Wellington
PSYC434: Conducting Research Across Cultures

Abstract

Background: Despite widespread concern about social media's impact on mental health, research yields inconsistent findings, potentially due to heterogeneous treatment effects that vary across individuals. Most studies assume uniform effects across users, which may obscure important individual differences. **Objectives**: 1. Estimate causal effects of high social media use (SMU) on psychological well-being. 2. Evaluate population variation in these effects. 3. Provide policy guidance on targeted interventions. **Method**: Longitudinal cohort study using the New Zealand Attitudes and Values Study (N=8,854 adults). We defined high SMU as ≥14 hours weekly and applied inverse probability weighting to address attrition and improve generalizability. Causal forests detected treatment effect heterogeneity, while policy trees identified individuals likely to benefit most from reduced SMU. **Results**: High SMU showed no significant average treatment effects on nine well-being outcomes. However, precision targeting strategies yielded meaningful benefits for body satisfaction and personal well-being in specific subgroups: individuals with low life meaning or liberal political orientation (body satisfaction), and heavy users with lower socioeconomic status (personal well-being). **Implications**: Results challenge assumptions of uniform social media effects and suggest personalised digital wellness strategies may be more effective than blanket policies. Mental health practitioners should consider individual characteristics when making social media recommendations rather than applying universal restrictions.

Keywords: *Social Media*; *Well-being, Causal Inference*; *Heterogeneous Treatment Effects*; *Longitudinal*; *Machine Learning*; *Policy Trees*; *Precision Medicine*.

While digital platforms offer unprecedented opportunities for social connection and information sharing, growing research has indicated that excessive use may compromise multiple dimensions of mental health, including anxiety, depression, rumination, perfectionism, low self-esteem, body dissatisfaction, and reduced personal well-being, life meaning, and life satisfaction. The scale of this concern becomes apparent when considering that global social media adoption has nearly doubled over the past decade, rising from 2.0 billion users in 2015 to over 5.0 billion by 2025 (Kemp, 2025), while in New Zealand, approximately 79% of the population, around 4.14 million people, are active social media users as of 2025 (Kemp, 2025). This widespread adoption makes understanding the mental health implications of social media use increasingly critical for public health policy and intervention development.

Depression and Anxiety

Studies have consistently associated general SMU, problematic use, and platform intensity with higher depression and anxiety levels. A systematic review of 32 studies found that over half reported positive associations between SMU and anxiety among adolescents, with problematic use measures showing the strongest associations, followed by screen time measures (Kerr et al., 2025). This pattern has been further supported by a meta-analysis of 18 studies that found moderate, statistically significant associations between problematic SMU and depression and anxiety in adolescents and young adults (Shannon et al., 2022).

Individual studies have demonstrated these associations across different measures of social media engagement. A study of young adults found that those in the highest quartile of SMU had significantly increased odds of depression compared to the lowest quartile users, with associations showing strong dose-response relationships across multiple measures of engagement (Lin, Sidani & Shensa, 2017). Users of 7-11 platforms show three times greater odds of high depressive and anxiety symptoms compared to those using ≤2 platforms (Primack et al., 2017). Problematic SMU, measured by addictive components including salience, mood modification, withdrawal, tolerance, and relapse, has been strongly associated with increased depressive symptoms, showing 9% increased odds among young adults (Shensa et al., 2017).

Rumination

Individuals prone to rumination are more susceptible to negative social media effects, while excessive use perpetuates ruminative thinking patterns, creating cycles that maintain psychological distress. In a study of adolescents, SMU was significantly associated with rumination, and rumination, in turn, mediated the relationship between online social comparison and depressive symptoms (Nesi & Prinstein, 2015). The mechanism of social comparison has been found to operate primarily through negative comparisons on platforms that increase rumination, which predicts higher depressive symptoms (Feinstein et al., 2013).

Further highlighting this relationship, a two-part study found that depressive symptoms among youth were linked to the quality rather than quantity of social networking interactions, with rumination and co-rumination both associated with negative experiences online (Davila

et al., 2021). For example, among young adults, depressive rumination moderated the impact of online interactions, suggesting that vulnerable individuals are more likely to interpret and engage with social media in ways that reinforce distress.

Perfectionism

Research has indicated that those with maladaptive evaluation concern (self-criticism) or adaptive achievement-striving (performance-based), both dimensions of perfectionism (Frost et al., 1993), distinct from appearance-oriented perfectionism, are more likely to engage in problematic SMU. In a study of Chinese college students, maladaptive perfectionism was directly and indirectly related to internet addiction, with depression as a key mediator (Yang et al., 2021). Similarly, perfectionism predicted a preference for online social interaction, which in turn predicted emotionally motivated use and poor control over social media habits (Fioravanti et al., 2020). Another study (Harren, 2021) found that social media burnout (exhaustion and distress from SMU) was significantly predicted by socially prescribed positive perfectionism (external pressure to be perfect) and self-oriented negative perfectionism (harsh self-criticism).

Body Satisfaction

Research demonstrates robust associations between SMU and compromised body satisfaction, with appearance-based social comparison serving as the primary underlying mechanism. Meta-analytic evidence has revealed strong correlations between social media appearance comparison and body image concerns across 83 studies (Bonfanti et al., 2025). A cross-sectional study found that higher daily Instagram use was associated with lower self-esteem, greater body dissatisfaction, and more frequent physical appearance comparisons, regardless of the type of content viewed (Alfonso-Fuertes et al., 2023). An experimental study found that brief exposure to Instagram images depicting hegemonic beauty ideals led to increased body dissatisfaction and lower mood, particularly among women, compared to exposure to body-diverse content **(**Castellanos Silva & Steins, 2023**).**

Personal Well-being and Life Satisfaction

Personal well-being is a multidimensional construct encompassing positive emotional states, effective psychological functioning, and the absence of negative affect, including stress and loneliness (Diener & Ryan, 2009). This concept integrates hedonic elements such as life satisfaction and positive mood with eudaimonic components including meaning and personal growth (Keyes, 2002). The relationship between SMU and well-being demonstrates age-related variation with mixed findings across populations. Among younger adults, research reveals conflicting results: some studies indicate that increased SMU correlates with diminished well-being, with both Facebook and Instagram use predicting reduced well-being through intermediate constructs such as self-esteem and repetitive negative thinking (Faelens et al., 2021). Conversely, other research suggests positive feedback mechanisms, where receiving Likes on social media platforms can enhance happiness through increased self-esteem (Marengo et al., 2021). For older adults, cross-sectional research suggests more positive associations, with greater social media engagement linked to reduced loneliness (Rennoch et al., 2023) and better mental health (Fu & Xie, 2021). However, problematic

SMU appears to affect well-being primarily through its impact on loneliness, with increased problematic use predicting greater loneliness over time, which subsequently reduces life satisfaction (Marttila et al., 2021).

Ethnic Disparities and Cultural Factors

Māori youth in Aotearoa New Zealand face a concerning digital landscape, experiencing disproportionate exposure to harmful online content including cyberbullying, hate speech, and self-harm imagery (Office of the Children's Commissioner, 2021). This digital vulnerability intersects with existing mental health disparities, as Māori report significantly higher rates of psychological distress and suicidal ideation compared to non-Māori populations, with mental distress rates nearly 50% higher (Government Inquiry into Mental Health and Addiction, 2018; Fleming et al., 2020).

The structural design of mainstream social media platforms presents additional challenges for Māori wellbeing. While research demonstrates that cultural identity, Te Reo Māori, and strong whakapapa-based relationships (whanaungatanga) serve as vital protective factors for Māori mental health (Rangihuna et al., 2020), Western-designed platforms predominantly emphasise individual self-presentation, personal achievement, and social comparison. This fundamental misalignment between platform design principles and Māori values of collectivism, community-centred identity, and holistic wellbeing may create additional psychological strain, potentially undermining the cultural connections that can protect Māori mental health.

Policy Context

There is consistent evidence that excessive SMU can undermine multiple aspects of psychological well-being. The World Health Organisation (WHO, 2023) has called for enhanced digital literacy, stricter age regulations, and the redesign of social media platforms to prioritise youth safety. In Aotearoa, health authorities have recommended limiting screen time, encouraging tech-free zones in households, and integrating digital literacy into school curricula (Ministry of Health, 2022). Recent public debates have also called for raising the minimum age of social media access from 13 to 16, following Australia's lead (Office of the Children's Commissioner, 2021).

The absence of Māori-specific research on the psychological impacts of social media represents a critical gap in the evidence base. Without culturally grounded insights, policies and interventions risk being ineffective or even harmful for Māori.

Our Study

Causal Structure

This study addresses the causal question: What is the effect of SMU on psychological well-being, and does this effect vary by ethnicity?

The following directed acyclic graph (Figure 1) illustrates our assumed causal relationships and confounding structure. We address unmeasured confounders through baseline measures

of SMU, well-being, and covariates, relying on three key identification assumptions: consistency, no unmeasured confounding, and positivity. The temporal ordering ensures proper causal identification, with all baseline confounders measured before the exposure period, and outcomes measured after exposure assessment. By conditioning on the complete set of baseline measures {SMU_0, WB_0, L_0}, we aim to block backdoor confounding paths and satisfy the conditional exchangeability assumption necessary for causal inference.

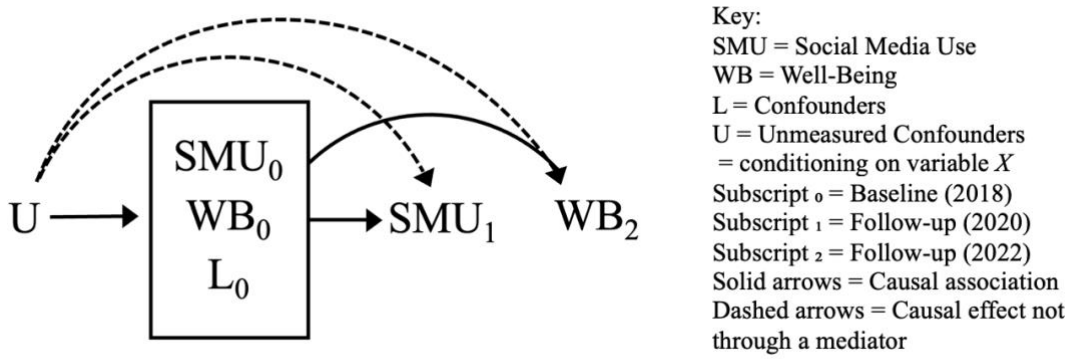


Figure 1: *Directed acyclic graph illustrating the causal structure between SMU and well-being outcomes.*

Intervention Framework and Study Rationale

From an intervention perspective, establishing meaningful usage thresholds represents a crucial step toward translating research findings into actionable recommendations. A threshold of 14 hours per week (approximately two hours daily) emerges as both empirically grounded and practically implementable across diverse intervention contexts.

We emulate a randomised controlled trial comparing high (≥14 hours/week) versus low (<14 hours/week) social media exposure. Valid causal inference requires three key assumptions: consistency (exposure captures meaningful intervention variation), no unmeasured confounding (all joint predictors of exposure and outcome are controlled), and positivity (non-zero probability of either exposure level given covariates).

This two-hour daily threshold offers multiple implementation pathways: platform-based interventions could trigger usage alerts or impose time limits at this level, public health campaigns could promote the "two-hour guideline" as a memorable health recommendation, and educational programs could establish this benchmark as a practical target for healthy digital habits. The threshold's specificity enables direct translation from research evidence to policy and clinical practice.

Current research limitations that our study addresses include: (1) limited causal evidence due to reliance on observational studies, (2) absence of ethnicity-specific research despite documented mental health disparities across ethnic groups, and (3) lack of policy-relevant usage thresholds necessary for evidence-based intervention development.

Methods

Study Design and Data Collection

This study employed a three-wave longitudinal design using data from the New Zealand Attitudes and Values Study (NZAVS), a national longitudinal probability study designed to reflect the adult New Zealand population. Data were collected at baseline (2018), during the exposure period (2020), and at outcome measurement (2022). This temporal structure was designed to support causal inference by clearly establishing the sequence of events: the 2018 wave captured initial social media use (SMU) patterns and mental health status, the 2020 wave assessed changes in SMU, and the 2022 wave allowed time for potential causal effects on psychological well-being to emerge.

Target Population and Eligibility Criteria

The target population comprised New Zealand residents aged 13 to 65 who participated in the NZAVS in 2018. While the NZAVS sampling frame aims to represent the adult New Zealand population, it systematically under-samples males and individuals of Asian descent while over-sampling females and Māori. To address these disparities, we applied New Zealand Census survey weights based on age, gender, and ethnicity to enhance representativeness (Sibley, 2021).

Participants were included if they: (1) participated in the 2018 wave of the NZAVS, (2) provided a valid response to the baseline measure of SMU in hours per week, (3) were aged between 13 and 65, and (4) reported using social media for more than one hour per week at baseline. The minimum usage threshold was applied because many participants reported never using social media, and including these non-users would have introduced significant heterogeneity. By focusing on participants who use social media at least somewhat regularly, the analysis compared high users (≥14 hours/week) to a meaningful lower-use reference group (1–13.9 hours/week).

A total of 8,854 individuals met these criteria and were included in the sample. The subgroup analysis included 7,392 NZ Europeans (83.49% of the sample) and 868 Māori participants (9.80% of the sample).

Exposure Definition

The NZAVS measured social media usage by asking participants to report the number of hours they spend using social media platforms per week, specifically phrased as "Hours spent using social media (e.g., Facebook)." Given the highly skewed distribution of reported hours, with most participants reporting low to moderate use and a small number reporting very high use, the natural logarithm (+1) was applied. This transformation reduced the influence of extreme values and stabilised the variance, thereby improving the suitability of the data for statistical analysis (Sibley et al., 2011) (see Appendix A).

SMU was operationalised as a binary exposure, with the split at 14 hours per week: high use (≥14 hours per week) versus lower use (<14 hours per week). This threshold represents approximately two hours of daily use, a point identified in existing research (Page et al.,

2010; Twenge et al., 2017; Twenge & Campbell, 2018) where adverse mental health outcomes become substantially more likely. The binary classification served multiple purposes: it aligns with natural breakpoints in the literature, creates actionable policy targets, and facilitates clear causal interpretation.

Outcome Measures

We examined psychological well-being across multiple dimensions, including positive indicators (life satisfaction, self-esteem, body satisfaction, life meaning) and negative indicators (depression, anxiety, rumination, perfectionism). This comprehensive approach acknowledged that social media may simultaneously harm some aspects of well-being while benefiting others. The breadth of measures allowed us to identify which specific dimensions are most vulnerable to high SMU and whether protective effects exist alongside harmful ones.

Our outcomes were continuous ordinal variables rather than binary measures. Depression and anxiety ranged from 0-4, life satisfaction and related measures spanned 1-7 scales, and the well-being index used 0-10 ratings. The continuous nature of these variables supported our analytical approach using linear models rather than logistic regression.

Baseline Confounders

Baseline confounders included demographic, socioeconomic, personality, health, and psychosocial variables measured before exposure assessment. These variables plausibly influence both social media usage patterns and well-being outcomes through shared risk pathways (e.g., neuroticism predicts both higher SMU and poorer mental health; socioeconomic status affects both digital access and health outcomes).

Our baseline confounders were: age, birth country, education, employment status, ethnicity, gender, sexual orientation, parental/partnership status, rurality, survey participation mode; personality traits (agreeableness, conscientiousness, neuroticism, openness); health and lifestyle factors (alcohol use, disability status, time allocations, household income, self-rated health, smoking); and social and psychological measures (belonging, deprivation index, socioeconomic status, political orientation, religiosity).

These confounders included binary variables (employment, gender, smoking), ordinal scales (education, political orientation, religiosity), and continuous measures (age, income, personality scores). Continuous variables were standardised to z-scores for comparable scaling, while categorical variables with three or more levels used one-hot encoding to preserve their categorical nature without imposing artificial ordering.

Causal Framework and Identification Assumptions

Target Trial Approach

We employed a target trial framework (Hernán et al., 2016) to clarify the causal question under investigation: "How would well-being outcomes change if, for everyone in the

population, we set social media exposure to ≥14 hours per week (approximately ≥2 hours daily), compared with setting it to <14 hours per week, given each individual's characteristics?"

This approach compared two counterfactual scenarios:

1. Treatment condition (A=1): Everyone receives high social media exposure (≥14 hours/week)
2. Control condition (A=0): Everyone receives low social media exposure (<14 hours/week)

The Average Treatment Effect (ATE) represents the population-level difference between these scenarios.

Identification Assumptions

This study relied on three key identification assumptions for estimating the causal effect of high SMU on well-being outcomes:

1. **Consistency**: The observed outcome under the observed SMU level equals the potential outcome under that exposure level. As part of consistency, we assumed no interference: the potential outcomes for one individual are not affected by the SMU status of other individuals.
2. **No unmeasured confounding**: All variables that affect both high SMU and well-being outcomes have been measured and accounted for in the analysis, including demographic characteristics, personality traits, and baseline psychological states.
3. **Positivity**: There is a non-zero probability of experiencing each level of SMU (high vs. low) for every combination of values of social media exposure and confounders in the population. Positivity is the only fundamental causal assumption that can be evaluated empirically with the available data (see Appendix C).

Confounding Control

To manage confounding, we implemented VanderWeele's (2019) modified disjunctive cause criterion by:

1. Identifying all common causes of SMU and well-being outcomes, including demographic factors, personality characteristics, and lifestyle variables.
2. Excluding instrumental variables that affect social media exposure but not well-being outcomes directly.
3. Including proxies for unmeasured confounders affecting both exposure and outcomes.
4. Controlling for baseline exposure and baseline outcomes as proxies for unmeasured common causes, enhancing the robustness of our causal estimates (VanderWeele et al., 2020).

Statistical Methods

Doubly Robust Estimation

We used doubly robust estimation for subgroup analysis, combining features of both inverse probability of treatment weighting (IPTW) and G-computation methods. This approach provides unbiased estimates if either the propensity score or outcome model is correctly specified. The process involved five main steps:

Step 1: Propensity Score Estimation - We estimated the conditional probability of exposure given covariates and subgroup indicators using logistic regression, with continuous predictors transformed to z-scores and categorical variables appropriately coded. Propensity scores were estimated separately within each ethnic subgroup stratum.

Step 2: Weighted Outcome Model - We fitted weighted outcome models using the calculated propensity score weights, estimating outcomes conditional on exposure, covariates, and subgroup.

Step 3: Potential Outcomes Simulation - We simulated potential outcomes for each individual in each subgroup under hypothetical scenarios of universal exposure to high versus low SMU.

Step 4: Average Causal Effect Estimation - We estimated the average causal effect for each subgroup by comparing the expected values of potential outcomes under each intervention level.

Step 5: Between-Group Comparisons - We calculated differences in estimated causal effects between NZ Europeans and Māori, with confidence intervals and standard errors determined using simulation-based inference methods via the clarify package in R (Greifer et al., 2023).

Treatment Effect Heterogeneity Analysis

We pursued two complementary objectives: testing whether personalised targeting based on individual conditional average treatment effects yields welfare gains, and translating such gains into transparent, practitioner-ready decision rules.

Pre-processing and Model Training: We used an honest 70/30 split where the training fold built the causal forest with grf (Tibshirani et al., 2024), while the held-out fold provided data for fitting policy trees. Outcomes were sign-flipped where necessary so that larger values always indexed improvement.

Budget-based Screening: We used Qini curves to quantify incremental gains when treating only the top 20% or 50% of individuals ranked by predicted benefit, compared to treating everyone uniformly. Outcomes whose 95% confidence intervals excluded zero at either spending level were labeled as "actionable."

Decision Rule Development: For each actionable outcome, we trained depth-2 policy trees with policytree (Athey & Wager, 2021a; Sverdrup et al., 2024) to create transparent if-then statements that maximize expected welfare.

Sensitivity Analysis

We performed sensitivity analyses using E-values (VanderWeele & Ding, 2017; Linden et al., 2020). E-values quantify the minimum strength of association that an unmeasured

confounder would need to have with both the exposure (SMU) and outcome (well-being measures) to fully explain away our observed effect estimates. Specifically, the E-value represents the minimum risk ratio that an unmeasured confounder would need to have with both SMU and each well-being outcome, conditional on our measured confounders, to reduce the observed effect to the null.

For each outcome measure, we calculated E-values using the formula: E-value = effect estimate + √(effect estimate × (effect estimate - 1)), where larger E-values indicate greater robustness to unmeasured confounding. We reported both the E-value for the point estimate and the lower confidence limit, with the latter providing a more conservative assessment of sensitivity to unmeasured confounding.

Confidence intervals for each E-value were derived from the multiplicity-adjusted confidence intervals of the corresponding coefficient estimates (Bonferroni correction, $\alpha = 0.05$), ensuring that the sensitivity analysis maintained the same error-control framework as the main results.

Limitations and Assumptions

Measurement Error

Social media usage was self-reported and may be subject to recall bias, social desirability bias, and imprecise estimation of time spent on platforms. Participants may underestimate usage due to automatic or habitual engagement or overestimate due to perceived negative consequences of high usage. Well-being measures, while validated, rely on subjective self-assessment and may be influenced by current mood states or response styles. These measurement errors could attenuate true associations (classical measurement error) or introduce bias if errors are systematic and related to both exposure and outcomes.

Missing Data

This analysis used synthetic data without missing values. In practice, missing data would require careful handling through multiple imputation or other appropriate methods to maintain the validity of causal inferences. Participants lost to follow-up in subsequent waves were adjusted using inverse probability of censoring weights.

Multiple Testing

Given multiple outcome measures, ATE confidence intervals were adjusted for multiplicity using Bonferroni correction at $\alpha = 0.05$.

Results

ATE Analysis

The ATE analysis examining the relationship between hours of SMU and multidimensional well-being outcomes yielded inconclusive results across all measured constructs (Figure 2). While point estimates suggested positive associations for anxiety, depression, rumination,

and perfectionism, and negative associations for life meaning, personal well-being index, body satisfaction, and life satisfaction, with self-esteem showing no association, the confidence intervals for all outcomes encompassed zero, indicating that none of these effects achieved statistical significance at conventional alpha levels. These null findings suggest insufficient evidence to establish reliable causal associations between social media usage duration and any of the examined well-being dimensions.

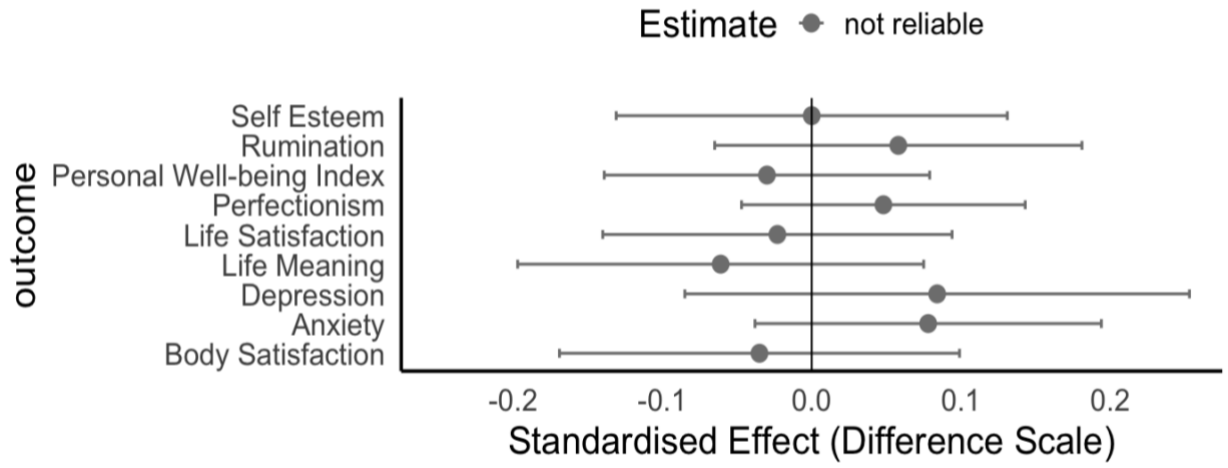


Figure 2*: Average treatment effects (ATEs) of SMU (in hours) on well-being.*

Table 1 presents the ATE calculations for each variable. Confidence intervals were adjusted for multiple comparisons using Bonferroni correction ($\alpha = 0.05$). E-values indicate the minimum strength of unmeasured confounding needed to nullify observed effects, which were also adjusted using Bonferroni correction ($\alpha = 0.05$). No outcomes showed reliable causal evidence (E-value lower bound > 1.2). All confidence intervals crossed zero after adjustment for multiple comparisons, indicating insufficient statistical evidence for causal effects of social media usage hours on any examined wellbeing dimensions. The E-value bounds of 1.0 suggest that the observed associations could be entirely explained by minimal unmeasured confounding.

Outcome	ATE	2.5 %	97.5 %	E-Value	E-Value Bound
Depression	0.084	-0.035	0.204	1.372	1
Anxiety	0.078	-0.004	0.160	1.355	1
Rumination	0.058	-0.029	0.145	1.293	1
Perfectionism	0.048	-0.019	0.115	1.261	1
Self Esteem	0.000	-0.092	0.093	1.000	1
Life Satisfaction	-0.023	-0.106	0.059	1.168	1
Personal Well-being Index	-0.030	-0.107	0.047	1.196	1
Body Satisfaction	-0.035	-0.129	0.060	1.215	1
Life Meaning	-0.061	-0.157	0.035	1.303	1

***Table 1**: Average Treatment Effects on Psychological Well-being Measures with 95% Confidence Intervals and E-values.*

Heterogeneous Treatment Effects

We begin by examining the distribution of individual treatment effects (τ_i) across our sample. Figure 3 presents the estimated treatment effects for each individual, revealing substantial variability in how people respond to lower SMU.

The substantial spread in each distribution indicates that reducing SMU does not affect everyone uniformly. For each outcome, some individuals would benefit substantially from lower SMU (positive τ_i values), others who might be harmed (negative τ_i values), and many who would experience minimal change.

To determine whether this variability is systematic (i.e., predictable based on individual characteristics) rather than random noise, we employ two complementary approaches: Qini curves to assess the reliability of heterogeneous effects, and policy trees to identify subgroups with differential treatment responses.

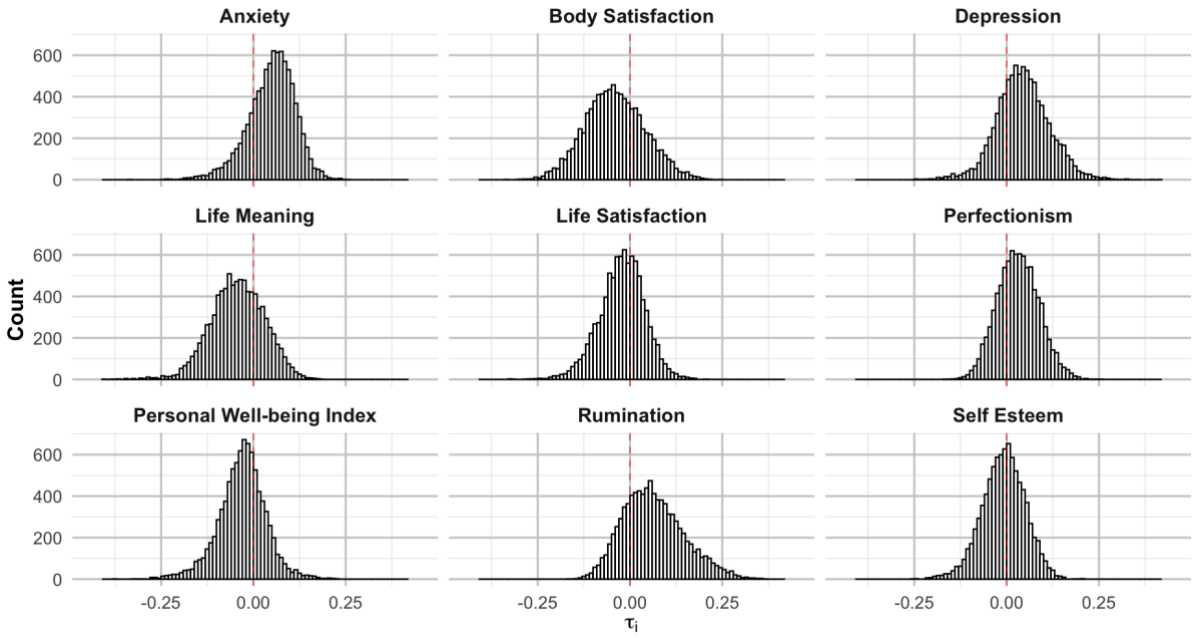


***Figure 3**: Distribution of Individual Treatment Effects (τ_i) Across Outcome.*

Qini Curves: Measuring Gains from Targeted Treatment at Different Budget Levels

The Qini curve shows the cumulative gain as we expand a targeting rule down the CATE ranking.

- Beneficial exposure: we add individuals from the top positive CATEs downward; the baseline is 'expose everyone.'
- Detrimental exposure: we first flip outcome direction (so higher values represent more harm; see Body Satisfaction, Self Esteem, Life Satisfaction, Personal Well-being Index, Life Meaning), then add the exposure starting with individuals whose CATEs show the greatest harm**,** gradually including those predicted to be more

resistant to harm; the baseline is 'expose everyone.' The curve, therefore, quantifies the harm when those most susceptible to harm are exposed.

If the Qini curve stays above its baseline, a targeted policy increases the outcome more than a one-size-fits-all alternative. Outcome directions were flipped where needed (Body Satisfaction, Life Meaning, Life Satisfaction, Personal Well-being Index, Self Esteem) so the positively valenced exposures always have positively valenced outcomes and negative exposures have negatively valenced outcomes.We computed cumulative gains from prioritising individuals by CATE at 20% and 50% spend levels, comparing against a no-prioritisation baseline (Table 2).

Body Satisfaction (reversed) At 20% spend: CATE prioritisation is beneficial (diff: 0.10 [95% CI 0.06, 0.14]). At 50% spend: CATE prioritisation is beneficial (diff: 0.10 [95% CI 0.07, 0.14]).
Anxiety At 20% spend: CATE prioritisation worsens outcomes compared to ATE. At 50% spend: CATE prioritisation worsens outcomes compared to ATE.
Depression No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.
Life Meaning (reversed) At 20% spend: CATE prioritisation worsens outcomes compared to ATE. At 50% spend: CATE prioritisation worsens outcomes compared to ATE.
Life Satisfaction (reversed) No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.
Perfectionism At 20% spend: CATE prioritisation worsens outcomes compared to ATE. At 50% spend: CATE prioritisation worsens outcomes compared to ATE.
Personal Well-being Index (reversed) At 20% spend: CATE prioritisation is beneficial (diff: 0.18 [95% CI 0.12, 0.25]). At 50% spend: CATE prioritisation is beneficial (diff: 0.19 [95% CI 0.10, 0.28]).
Rumination At 20% spend: CATE prioritisation worsens outcomes compared to ATE. At 50% spend: CATE prioritisation worsens outcomes compared to ATE.
Self Esteem (reversed) No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.

Model	Spend 20%	Spend 50%
Body Satisfaction (reversed)	**0.10 [0.06, 0.14]**	**0.10 [0.07, 0.14]**
Anxiety	*-0.07 [-0.12, -0.02]*	*-0.14 [-0.22, -0.06]*
Depression	-0.05 [-0.11, 0.00]	-0.05 [-0.13, 0.04]
Life Meaning (reversed)	*-0.07 [-0.10, -0.05]*	*-0.07 [-0.10, -0.05]*
Life Satisfaction (reversed)	-0.02 [-0.05, 0.00]	-0.02 [-0.05, 0.00]
Perfectionism	*-0.11 [-0.16, -0.06]*	*-0.20 [-0.28, -0.13]*
Personal Well-being Index (reversed)	**0.18 [0.12, 0.25]**	**0.19 [0.10, 0.28]**
Rumination	*-0.17 [-0.22, -0.13]*	*-0.15 [-0.21, -0.09]*
Self Esteem (reversed)	0.06 [-0.00, 0.12]	0.04 [-0.04, 0.12]

Table 2: *Qini Curve Results showing estimated effects and 95% confidence intervals for each outcome at 20% and 50% spend levels. Bold values indicate beneficial effects of CATE prioritisation. Italicised values indicate harmful effects where CATE prioritisation worsens outcomes compared to average treatment effects (ATE).*

Figure 4 presents the results for reliable Qini results.

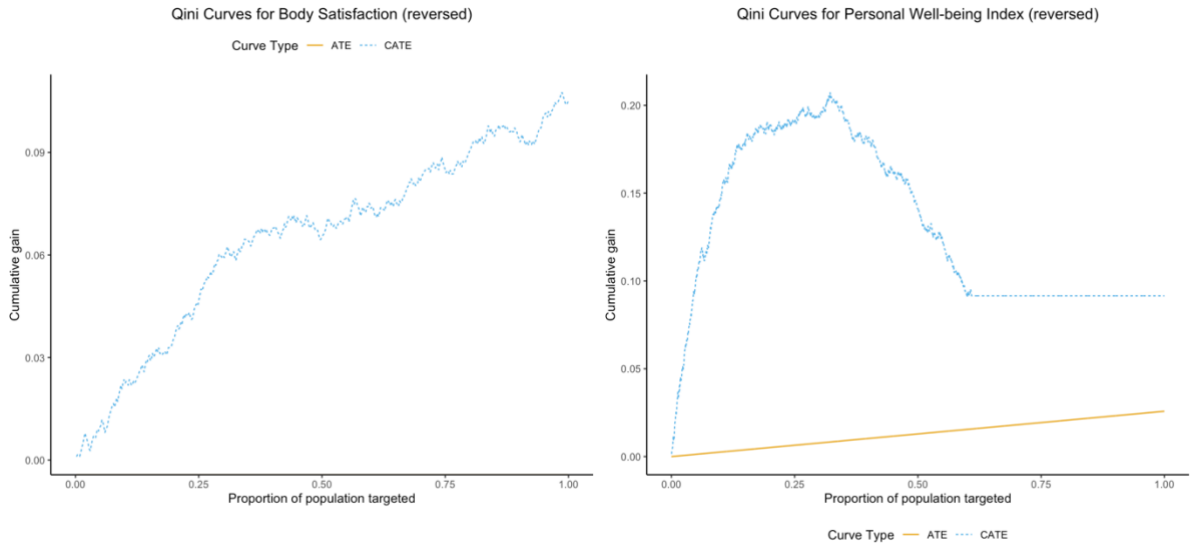


Figure 4*: Qini Graphs for Body Satisfaction and Personal Well-Being*

Interpretation of Qini Curves

Body Satisfaction The Qini curve for Body Satisfaction demonstrates strong evidence for the effectiveness of individualised treatment targeting. The CATE-based prioritisation curve shows a steady, monotonic increase in cumulative gain as the treatment budget expands, reaching approximately 0.09 units by full population coverage. This consistent upward trajectory indicates that targeting individuals based on their predicted treatment response delivers significantly greater benefits than uniform treatment allocation. The smooth, continuous gains suggest that heterogeneous treatment effects exist across the population, with the CATE model successfully rank-ordering individuals from highest to lowest treatment responsiveness. The substantial gap between the CATE curve and what would be expected from random allocation demonstrates the model's ability to identify meaningful differences in individual treatment effects.

Personal Well-being Index The Personal Well-being Index shows a more pronounced pattern of targeting benefits, with the CATE curve exhibiting a distinctive peak-decline-plateau shape. The curve rises sharply to reach maximum gains of approximately 0.20 units when treating around 25-30% of the population, then declines slightly before flattening into a plateau as the budget expands to include the full population. This pattern suggests that the intervention has the strongest positive effects on a specific subset of individuals, followed by a population with modest negative treatment effects, and finally a large group of non-responders. The ATE baseline (orange line) remains relatively flat near zero, highlighting that

uniform treatment would provide minimal population-level benefit, while targeted treatment delivers substantial improvements for those most likely to respond positively, while avoiding both non-responders and those who may experience small negative effects.

Policy Trees

We used policy trees (Athey & Wager, 2021b, 2021a; Sverdrup et al., 2024) to find straightforward 'if-then' rules for who benefits most from treatment, based on participant characteristics. Because we flipped some measures, a higher predicted effect always means greater improvement. Policy trees can uncover small but important subgroups whose treatment responses stand out, even when the overall differences might be modest.

A shallow policy tree recommends actions based on two splits for depth=2, or one split for depth=1. We trained on 50% of the data and evaluated on the rest. Each tree shows: (1) the decision rules for treatment assignment, (2) the distribution of treatment effects across subgroups, and (3) a visual representation of how covariates split the population into groups with differential treatment responses.

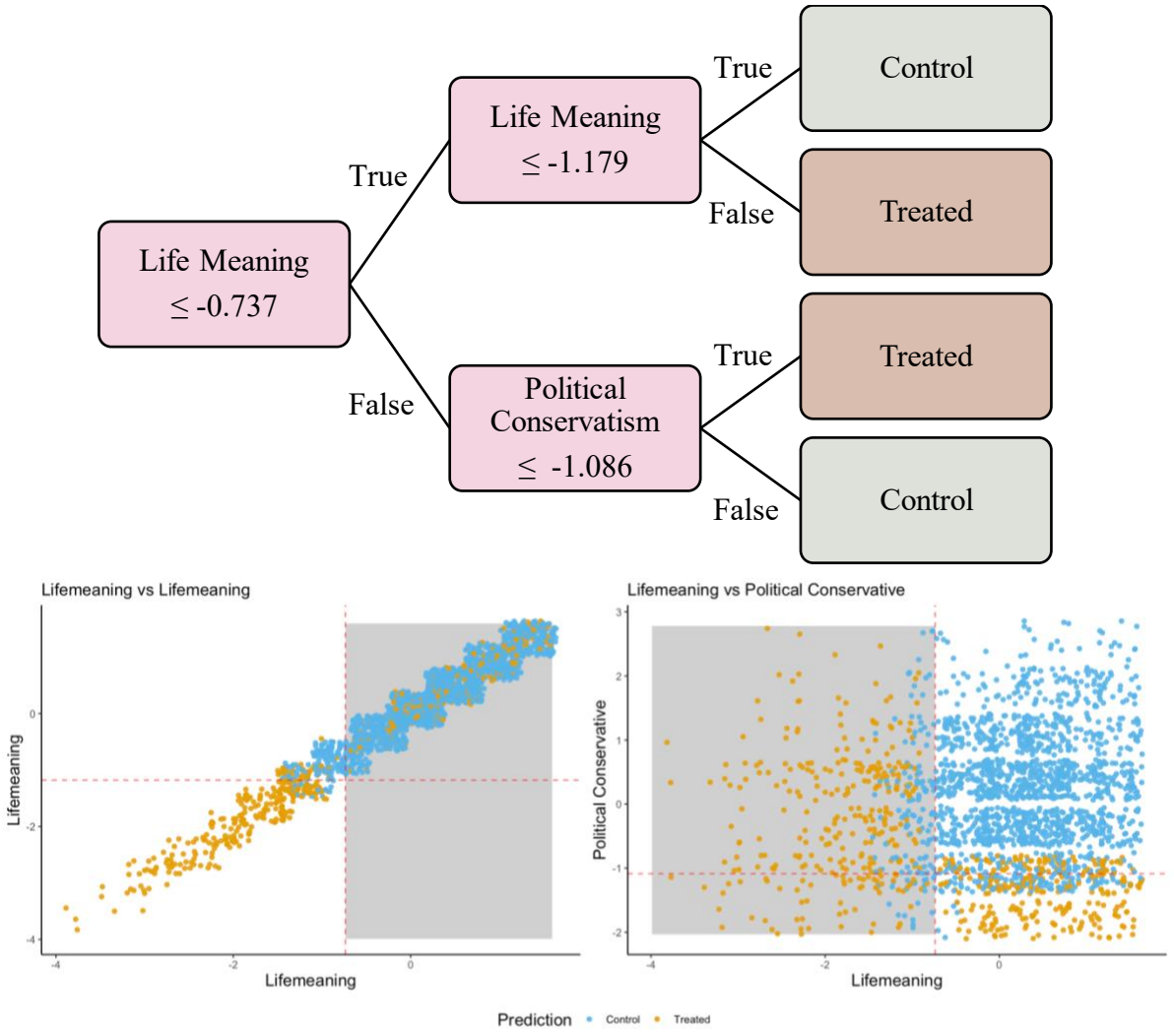


Figure 5*: Policy Tree for Body Satisfaction. The top panel shows the decision tree for optimal allocation of social media reduction treatment based on life meaning and political*

conservatism. The bottom left panel shows the primary split based on Life Meaning (vertical dashed line at -0.737) and secondary split for the lower life meaning group (horizontal dashed line at -1.179). The bottom right panel displays the decision boundary for the higher life meaning group based on Political Conservative scores (horizontal dashed line at -1.086). Orange points represent individuals assigned to the social media reduction treatment, while blue points represent those assigned to control.

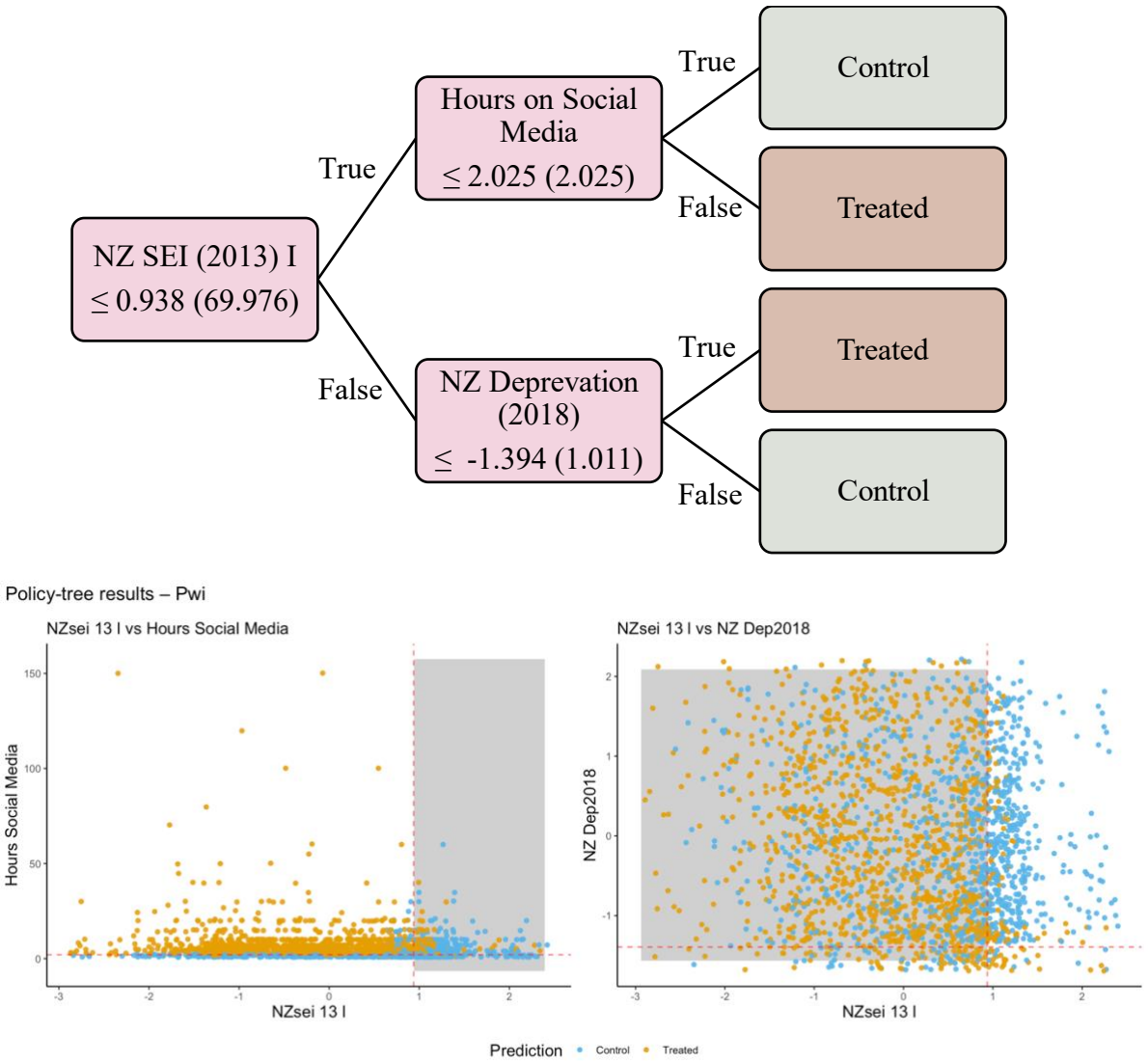


***Figure 6**: Policy Tree for Body Satisfaction. The top panel shows the decision tree for optimal allocation of social media reduction treatment based on socioeconomic status, social media usage, and area deprivation. The bottom left panel shows the decision boundary for the lower socioeconomic group (NZsei 13 ≤ 0.938) based on Hours Social Media usage (horizontal dashed line at 2.025 hours). The bottom right panel displays the decision boundary for the higher socioeconomic group (NZsei 13 > 0.938) based on NZ Dep2018 area deprivation scores (horizontal dashed line at -1.394). Orange points represent individuals assigned to the social media reduction treatment, while blue points represent those assigned to control.*

Policy Tree Interpretations (depth 2)

Findings for Body Satisfaction:

Split 1: Life Meaning ≤ -0.737. Within that subgroup, split 2a: Life Meaning ≤ -1.179, → **Control**; Life Meaning ≤ -1.179 → **Treated**.

Split 2: Life Meaning ≤ -0.737. Within that subgroup, split 2b: Political Conservative ≤ -1.086, → **Treated**; Political Conservative ≤ -1.086 → **Control**.

Findings for Personal Well-being Index:

Split 1: NZ Socio-Economic Index 2013 1 ≤ 0.938 (original: 69.976). Within that subgroup, split 2a: Hours of Social Media ≤ 2.025 (original: 2.025), → **Control**; Hours of Social Media ≤ 2.025 (original: 2.025) → **Treated**.

Split 2: NZ Socio-Economic Index 2013 1 ≤ 0.938 (original: 69.976). Within that subgroup, split 2b: NZ Deprivation 2018 ≤ -1.394 (original: 1.011), → **Treated**; NZ Deprivation 2018 ≤ -1.394 (original: 1.011) → **Control**.

Subgroup Analysis by Ethnicity

The subgroup analysis examining treatment effects across ethnic groups revealed no statistically significant differences in how social media usage affects well-being outcomes between NZ Europeans and Māori populations (Figure 8).

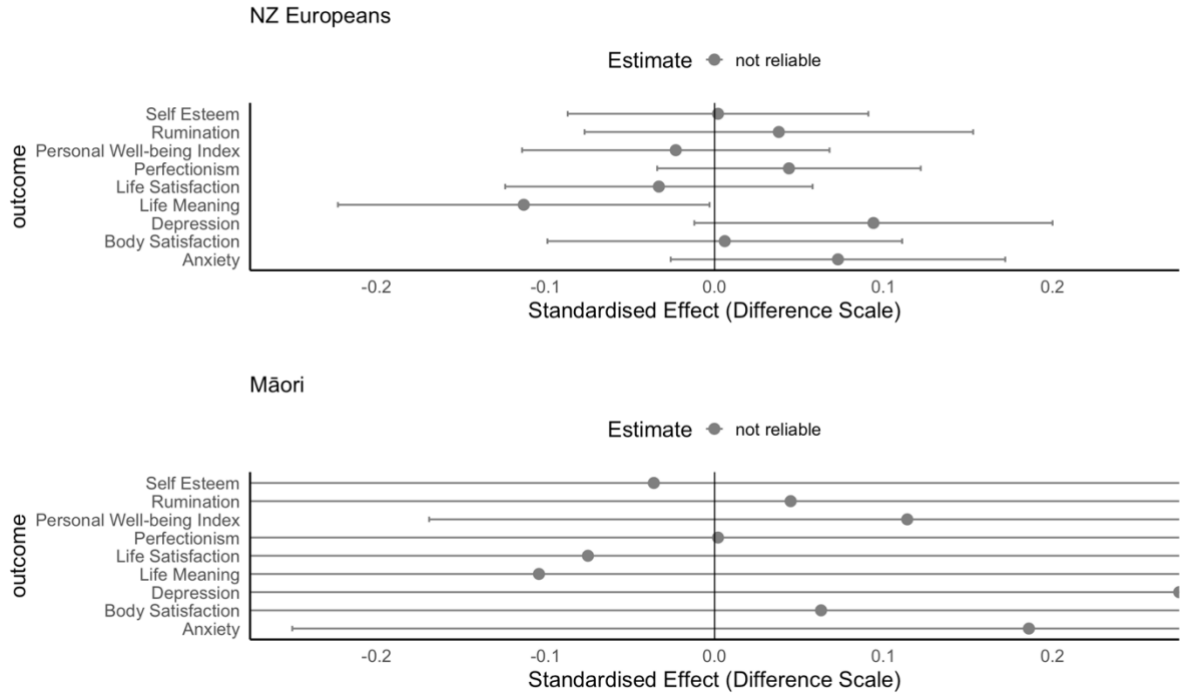


Figure 2*: Standardized ATE of social media on well-being outcomes for NZ Europeans (top) and Māori (bottom).*

For NZ Europeans, only Life Meaning showed a statistically significant effect, with a negative association, as indicated by the confidence interval not crossing zero. This indicates

that increased social media usage is associated with decreased life meaning in this population. All other outcomes (self-esteem, rumination, personal well-being index, perfectionism, life satisfaction, depression, body satisfaction, and anxiety) showed non-significant effects with confidence intervals encompassing zero.

For Māori participants, no outcomes showed statistically significant effects. All confidence intervals crossed zero, indicating insufficient evidence for reliable causal effects of social media usage on any examined well-being dimension in this population. Notably, the confidence intervals for the Māori subgroup are substantially wider than those for NZ Europeans across all outcomes, reflecting the much smaller sample size and resulting in considerably less precision in the effect estimates.

Confidence intervals were adjusted for multiple comparisons using Bonferroni correction ($\alpha = 0.05$). No outcomes showed reliable evidence of treatment effect heterogeneity between ethnic groups, as all confidence intervals encompassed zero. The substantially wider confidence intervals observed in the individual forest plots for Māori participants reflect the smaller sample size, but the group difference analysis indicates that social media effects on well-being do not differ significantly between NZ Europeans and Māori populations. These findings suggest homogeneous treatment effects across ethnic groups.

Outcome	Group Difference	95% CI
Depression	0.181	[-0.222, 0.584]
Personal Well-being Index	0.137	[-0.073, 0.347]
Anxiety	0.113	[-0.203, 0.429]
Body Satisfaction	0.057	[-0.251, 0.365]
Life Meaning	0.009	[-0.316, 0.334]
Rumination	0.007	[-0.276, 0.290]
Self Esteem	-0.038	[-0.374, 0.298]
Perfectionism	-0.042	[-0.255, 0.171]
Life Satisfaction	-0.042	[-0.423, 0.339]

Table 3*: Group Differences in Treatment Effects (Māori vs NZ Europeans)*

Discussion

Principal Findings

This longitudinal causal analysis found no significant population-level effects of high SMU (≥14 hours/week) on psychological well-being across nine outcomes after Bonferroni corrections. While point estimates suggested negative trends, worse scores on anxiety, depression, rumination, and perfectionism, and lower scores on life meaning, satisfaction, and well-being, all confidence intervals included zero, indicating no reliable causal effects.

However, heterogeneous treatment effects showed substantial individual variation. Qini curve analysis highlighted potential benefits of targeted interventions for body satisfaction and personal well-being. Policy tree analysis identified subgroups likely to benefit from reduced SMU, suggesting that tailored approaches may be more effective than broad ones.

The ethnicity-focused analysis found no significant differences in treatment effects between NZ Europeans and Māori, suggesting treatment effects were broadly similar across groups. This challenges my initial predictions that we would find greater mental health vulnerability among Māori.

Interpretation of E-values

The E-values calculated for this study indicate minimal robustness to unmeasured confounding, with all E-value bounds equal to 1.0 and point estimates ranging from 1.0 to 1.372. These values represent the minimum strength of association (on the risk-ratio scale) that an unmeasured confounder would need with both social media exposure and well-being outcomes to fully explain away the observed associations.

The E-value bounds of 1.0 are particularly concerning as they suggest that even minimal unmeasured confounding could entirely explain the observed associations. For context, an E-value of 1.2 is often considered a threshold for meaningful robustness, as it would require an unmeasured confounder with at least a 20% association with both exposure and outcome. The failure to exceed this threshold across all outcomes indicates that the observed associations are highly vulnerable to residual confounding.

This sensitivity analysis suggests that despite comprehensive baseline confounder control, including demographic, personality, socioeconomic, and psychosocial variables, important unmeasured confounders may still influence the relationship between SMU and well-being. Potential candidates include genetic predispositions to mental health conditions, unmeasured social environmental factors, or time-varying confounders not captured in our three-wave design.

Causal Effect Interpretation

While no population-level causal effects were statistically significant, the heterogeneous treatment effects analysis revealed meaningful variation in individual responses to social media reduction. The wide distribution of individual treatment effects (τ_i) suggests some individuals benefit, others may be harmed, and many show minimal change.

For outcomes with reliable targeting benefits, effect sizes were practically significant. Body satisfaction improved by ~0.10 units and personal well-being by 0.18–0.19 units through CATE prioritisation. The policy tree analysis offered actionable insights. For body satisfaction, the most responsive groups were those with moderate to high life meaning, especially if less politically conservative. For personal well-being, benefits emerged among lower-SES individuals depending on usage level, and higher-SES individuals in less deprived areas.

Uncertainty and Confidence Intervals
The wide confidence intervals across most outcomes, especially after Bonferroni correction, highlight substantial uncertainty in the effect estimates and limit the strength of population-level conclusions.

In the ATE analysis, even the largest point estimates (e.g., 0.084 for depression) had confidence intervals spanning both beneficial and harmful effects, making clear recommendations difficult. In contrast, the heterogeneous treatment effects analysis showed greater precision. For example, Qini curve confidence intervals for body satisfaction (0.06–0.14 at 20% spend) were narrower, supporting the value of targeted over universal interventions, consistent with precision medicine principles.

The smaller Māori sample resulted in wider confidence intervals and lower statistical power, limiting the ability to detect true differences. These wide intervals suggest that meaningful effects may exist but were not observable in this study.

Generalisability and Transportability
The 14-hour weekly threshold used in this study may not generalise across cultural or technological contexts. Social media usage norms, platform preferences, and the role of digital engagement differ widely across countries. While the two-hour daily benchmark is common in Western research, it may carry different meanings in societies where social media serves distinct social, economic, or cultural functions.

Temporal transportability also limits generalisability. The data reflect usage patterns from around 2020, a period shaped by the COVID-19 pandemic, which significantly altered digital behaviours. Since then, social media environments have shifted: platforms like TikTok have popularised algorithm-driven content over social networking, and public discourse around digital wellness has intensified. These changes, alongside advances in algorithmic targeting and evolving user habits, mean that the psychological impacts of social media use today may differ substantially from those observed in this dataset.

Assumptions and Limitations
This study relied on three key causal identification assumptions, each with notable limitations:

First, the consistency assumption requires that observed outcomes under observed SMU match potential outcomes under that exposure. However, "social media use" is a broad, heterogeneous category. Someone spending 14 hours a week browsing Instagram may have vastly different experiences than someone using the same time for Facebook group discussions or Twitter news.

Second, the no unmeasured confounding assumption, while addressed through extensive covariate control, remains difficult in social media research. E-values indicate that even minimal unmeasured confounding, such as genetic predispositions to mental health issues,

peer norms, algorithmic exposure, or time-specific stress, could explain the observed associations.

Third, the positivity assumption, which requires a non-zero probability of each exposure level given covariates, may not hold for all subgroups. Some individuals with specific demographic or psychological profiles may have near-zero probability of sustained high SMU, while others may find sustained low use nearly impossible given their social or occupational circumstances.

Theoretical Relevance

These findings have important implications for digital media effects theories. The null population-level effects challenge simple linear models of social media harm common in public discourse and some academic work. Instead, results support nuanced approaches emphasising individual differences and context. The heterogeneous treatment effects align with the differential susceptibility model, which posits media effects vary by individual traits rather than uniformly. Identifying life meaning, political orientation, and socioeconomic status as moderators supports frameworks highlighting psychological resources and social context as protective or risk factors.

The lack of significant ethnic differences challenges cultural-deficit models assuming greater indigenous vulnerability to digital harms. However, this should be interpreted cautiously due to measurement and sample size limits. It may instead support universalist approaches to digital wellness while recognising the need for culturally responsive interventions.

Policy tree findings suggest targeting based on psychological state (life meaning) and structural factors (socioeconomic status, deprivation) may be more effective than demographic targeting, supporting theories emphasising proximal psychological processes over distal demographics in media effects.

Implications for Public Policy and Platform Design

The heterogeneous treatment effects observed in this study have important implications for evidence-based policy development. Current debates about social media regulation often assume uniform effects across populations, but our findings suggest that one-size-fits-all approaches may be suboptimal. Instead, policy frameworks might benefit from incorporating individual differences and allowing for personalised approaches to digital well-being.

For digital platforms, these findings suggest potential value in developing personalised well-being features that consider individual user characteristics. However, such approaches raise significant ethical questions about surveillance, autonomy, and the potential for manipulative design. The decision rules identified in our policy trees could inform the development of optional, user-controlled features that help individuals make informed decisions about their social media engagement.

The minimal robustness to unmeasured confounding (low E-values) suggests that policy decisions should remain cautious about causal claims from observational research. While this study provides the strongest available causal evidence using longitudinal observational data, randomised controlled trials remain necessary for definitive causal conclusions about social media interventions.

Replication and Future Research

Future research should focus on several methodological improvements to better understand social media's causal effects on well-being. Larger samples, especially among ethnic minorities, are needed to improve power for detecting subgroup differences and reduce estimate uncertainty.

Improved exposure measurement is essential to address heterogeneity. Studies should examine platform-specific effects, content types (e.g., social comparison vs. news), active versus passive use, and algorithm-driven exposure. Ecological momentary assessment can capture real-time usage and immediate psychological responses, clarifying causal mechanisms. Longer follow-up with more frequent measurements would better capture dynamic and adaptation effects, as some mental health impacts may emerge or diminish over time. Cross-cultural replication, especially with indigenous populations using culturally appropriate methods and frameworks, is critical rather than applying Western measures universally.

Finally, the heterogeneous treatment effects approach should be expanded. Future work should test whether moderators like life meaning, political orientation, and socioeconomic status hold across populations and time. Machine learning may also reveal novel moderators beyond traditional measures.

Conclusions

This study provides evidence that the relationship between SMU and well-being is more complex and individualised than commonly assumed. While we found no evidence for universal effects of social media exposure duration on well-being outcomes, the presence of meaningful heterogeneous treatment effects suggests that personalised approaches to digital well-being may be both necessary and beneficial. These findings challenge simplistic narratives about social media's impact while pointing toward more nuanced, evidence-based approaches to supporting psychological well-being in digital environments.

References

Alfonso-Fuertes, I., Alvarez-Mon, M. A., Sanchez-del-Hoyo, R., Ortega, M. A., Alvarez-Mon, M., & Molina-Ruiz, R. M. (2022). The time spent in Instagram is associated with greater dissatisfaction with body image, lower self-esteem and greater tendency to physical comparison among young adults in Spain: an Observational Study. *JMIR Formative Research*, *7*(1). https://doi.org/10.2196/42207

Ataera-Minster, J., & Trowland, H. (2024). Te Kaveinga: Mental health and wellbeing of Pacific peoples. *Results from the New Zealand Mental Health Monitor & Health and Lifestyles Survey*. Wellington: Health Promotion Agency. https://doi.org/10.60967/u002Fhealthnz.26536384.v1

Athey, S., & Wager, S. (2021a). Policy Learning With Observational Data. *Econometrica*, *89*(1), 133–161. https://doi.org/10.3982/ecta15732

Athey, S., & Wager, S. (2021b). Policy Learning With Observational Data. *Econometrica*, *89*(1), 133–161. https://doi.org/10.3982/ecta15732

Atkinson, J., Salmond, C., & Crampton, P. (2019). *NZDep2018 Index of Deprivation User's Manual*. https://www.otago.ac.nz/__data/assets/pdf_file/0030/314976/nzdep2018-index-of-deprivation-users-manual-730391.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bonfanti, R. C., Melchiori, F., Teti, A., Albano, G., Raffard, S., Rodgers, R., & Lo Coco, G. (2024). The association between social comparison in social media, body image concerns and eating disorder symptoms: A systematic review and meta-analysis. *Body Image*, *52*(101841), 101841. https://doi.org/10.1016/j.bodyim.2024.101841

Bulbulia, J. A. (2024a). *Margot: MARGinal observational treatment-effects*. https://doi.org/10.5281/zenodo.10907724

Bulbulia, J. A. (2024). Methods in Causal Inference Part 3: Measurement Error and External Validity Threats. *Evolutionary Human Sciences*, *6*(42). https://doi.org/10.1017/ehs.2024.33

Cummins, R., Eckersley, R., Pallant, J., Van Vugt, J., & Misajon, R. (2003). *Developing a national index of subjective wellbeing: The Australian Unity Wellbeing Index*.

Cutrona, C. E., & Russell, D. W. (1987). The provisions of social relationships and adaptation to stress. *Advances in Personal Relationships*, *1*, 37–67.

Davila, J., Hershenberg, R., Feinstein, B. A., Gorman, K., Bhatia, V., & Starr, L. R. (2012). Frequency and quality of social networking among young adults: Associations with depressive symptoms, rumination, and corumination. *Psychology of Popular Media Culture*, *1*(2), 72–86. https://doi.org/10.1037/a0027512

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*(1), 71–75.

EPIC Lab (Bulbulia). (2025). *Data templates and simulated data from the New Zealand Attitudes and Values Study.*

Faelens, L., Hoorelbeke, K., Soenens, B., Van Gaeveren, K., De Marez, L., De Raedt, R., & Koster, E. H. W. (2021). Social media use and well-being: A prospective experience-

sampling study. *Computers in Human Behavior*, *114*(106510), 106510. https://doi.org/10.1016/j.chb.2020.106510

Fahy, K. M., Lee, A., & Milne, B. J. (2017). *New Zealand socio-economic index 2013*. Statistics New Zealand-Tatauranga Aotearoa.

Feinstein, B. A., Hershenberg, R., Bhatia, V., Latack, J. A., Meuwly, N., & Davila, J. (2013). Negative social comparison on Facebook and depressive symptoms: Rumination as a mechanism. *Psychology of Popular Media Culture*, *2*(3), 161–170. https://doi.org/10.1037/a0033111

Fioravanti, G., Flett, G., Hewitt, P., Rugai, L., & Casale, S. (2020). How Maladaptive Cognitions Contribute to the Development of Problematic Social Media Use. *Addictive Behaviors Reports*, *11*(100267). https://doi.org/10.1016/j.abrep.2020.100267

Fleming, T., Crengle, S., Peiris-John, R., Ball, J., Fortune, S., Yao, E. S., Veukiso-Ulugia, A., & Clark, T. C. (2024). Priority actions for improving population youth mental health: An equity framework for Aotearoa New Zealand. *Mental Health & Prevention*, *34*(200340). https://doi.org/10.1016/j.mhp.2024.200340

Fraser, G., Bulbulia, J. A., Greaves, L., Wilson, M. S., & Sibley, C. G. (2020). Coding responses to an open-ended gender measure in a New Zealand national sample. Journal of Sex Research, 57, 979-986.

Fraser, G., Bulbulia, J., Greaves, L. M., Wilson, M. S., & Sibley, C. G. (2019). Coding Responses to an Open-ended Gender Measure in a New Zealand National Sample. *The Journal of Sex Research*, *57*(8), 979–986. https://doi.org/10.1080/00224499.2019.1687640

Frost, R. O., Heimberg, R. G., Holt, C. S., Mattia, J. I., & Neubauer, A. L. (1993). A comparison of two measures of perfectionism. *Personality and Individual Differences*, *14*(1), 119–126. https://doi.org/10.1016/0191-8869(93)90181-2

Fu, L., & Xie, Y. (2021). The Effects of Social Media Use on the Health of Older Adults: An Empirical Analysis Based on 2017 Chinese General Social Survey. *Healthcare*, *9*(9), 1143. https://doi.org/10.3390/healthcare9091143

Greaves, L. M., Barlow, F. K., Lee, C. H. J., Matika, C. M., Wang, W., Lindsay, C.-J., Case, C. J. B., Sengupta, N. K., Huang, Y., Cowie, L. J., Stronge, S., Storey, M., De Souza, L., Manuela, S., Hammond, M. D., Milojev, P., Townrow, C. S., Muriwai, E., Satherley, N., & Fraser, G. (2016). The Diversity and Prevalence of Sexual Orientation Self-Labels in a New Zealand National Sample. *Archives of Sexual Behavior*, *46*(5), 1325–1336. https://doi.org/10.1007/s10508-016-0857-5

Greaves, L. M., Barlow, F. K., Lee, C. H., Matika, C. M., Wang, W., Lindsay, C.-J., Case, C. J., Sengupta, N. K., Huang, Y., Cowie, L. J., et al. (2017). The diversity and prevalence of sexual orientation self-labels in a New Zealand national sample. *Archives of Sexual Behavior*, *46*, 1325–1336.

Hagerty, B. M. K., & Patusky, K. (1995). Developing a measure of sense of belonging. *Nursing Research*, *44*(1), 9–13. https://doi.org/10.1097/00006199-199501000-00003

Harren, N., Walburg, V., & Chabrol, H. (2021). Studying Social Media Burnout and Problematic Social Media use: The implication of perfectionism and metacognitions.

Computers in Human Behavior Reports, *4*(100117), 100117. https://doi.org/10.1016/j.chbr.2021.100117

Health, Ministry of. (2013). *The New Zealand Health Survey: Content guide 2012-2013*. Princeton University Press.

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, *79*, 70–75.

Instrument Ware Jr, J., & Sherbourne, C. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.

Jost, J. T. (2006). The end of the end of ideology. *American Psychologist*, *61*(7), 651–670. https://doi.org/10.1037/0003-066X.61.7.651

Karim, F. (2020). Social Media Use and Its Connection to Mental Health: A Systematic Review. *Cureus*, *12*(6). National Library of Medicine. https://doi.org/10.7759/cureus.8627

Kemp, S. (2025). *Global Social Media Statistics*. DataReportal – Global Digital Insights; Kepios. https://datareportal.com/social-media-users

Kemp, S. (2025). *Digital 2025: New Zealand*. DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2025-new-zealand

Kerr, B., Garimella, A., Pillarisetti, L., Charlly, N., Sullivan, K., & Moreno, M. A. (2024). Associations Between Social Media Use and Anxiety Among Adolescents: A Systematic Review Study. *Journal of Adolescent Health*, *76*(1). https://doi.org/10.1016/j.jadohealth.2024.09.003

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Keyes, C. L. M. (2002). The Mental Health Continuum: From Languishing to Flourishing in Life. *Journal of Health and Social Behavior*, *43*(2), 207–222. https://doi.org/10.2307/3090197

Linden, A., Mathur, M. B., & VanderWeele, T. J. (2020). Conducting sensitivity analysis for unmeasured confounding in observational studies using e-values: The evalue package. *The Stata Journal*, *20*(1), 162–175.

Lin, L. yi, Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., Hoffman, B. L., Giles, L. M., & Primack, B. A. (2016). Association between Social Media Use and Depression among US Young Adults. *Depression and Anxiety*, *33*(4), 323–331. https://doi.org/10.1002/da.22466

Marengo, D., Montag, C., Sindermann, C., Elhai, J. D., & Settanni, M. (2021). Examining the links between active Facebook use, received likes, self-esteem and happiness: A study using objective social media data. *Telematics and Informatics*, *58*(101523), 101523. https://doi.org/10.1016/j.tele.2020.101523

Marttila, E., Koivula, A., & Räsänen, P. (2021). Does Excessive Social Media Use Decrease Subjective well-being? a Longitudinal Analysis of the Relationship between

Problematic use, Loneliness and Life Satisfaction. *Telematics and Informatics*, *59*, 101556. https://doi.org/10.1016/j.tele.2020.101556

McComb, S. E., & Mills, J. S. (2021). Young women's body image following upwards comparison to Instagram models: The role of physical appearance perfectionism and cognitive emotion regulation. *Body Image*, *38*, 49–62. https://doi.org/10.1016/j.bodyim.2021.03.012

Nesi, J., & Prinstein, M. J. (2015). Using Social Media for Social Comparison and Feedback-Seeking: Gender and Popularity Moderate Associations with Depressive Symptoms. *Journal of Abnormal Child Psychology*, *43*(8), 1427–1438. https://doi.org/10.1007/s10802-015-0020-0

New Zealand Government. (2018). *He Ara Oranga: Report of the Government Inquiry into Mental Health and Addiction*. https://mentalhealth.inquiry.govt.nz/inquiry-report/he-ara-oranga

Nolen-hoeksema, S., & Morrow, J. (1993). Effects of rumination and distraction on naturally occurring depressed mood. *Cognition and Emotion*, *7*(6), 561–570. https://doi.org/10.1080/02699939308409206

Page, A. S., Cooper, A. R., Griew, P., & Jago, R. (2010). Children's Screen Viewing is Related to Psychological Difficulties Irrespective of Physical Activity. *Pediatrics*, *126*(5), e1011–e1017. https://doi.org/10.1542/peds.2010-1154

Pearl, J. (2009). *Causality*. Cambridge University Press.

Primack, B. A., Shensa, A., Escobar-Viera, C. G., Barrett, E. L., Sidani, J. E., Colditz, J. B., & James, A. E. (2017). Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among U.S. young adults. *Computers in Human Behavior*, *69*(1), 1–9. https://doi.org/10.1016/j.chb.2016.11.013

Rangihuna, D., Kopua, M., & Tipene-Leach, D. (2018). Mahi a Atua: a pathway forward for Māori mental health? *The New Zealand Medical Journal*, *131*(1471), 79–83.

Rennoch, G., Schlomann, A., & Zank, S. (2023). The Relationship Between Internet Use for Social Purposes, Loneliness, and Depressive Symptoms Among the Oldest Old. *Research on Aging*, *45*(9). https://doi.org/10.1177/01640275221150017

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Sengupta, N. K., Luyten, N., Greaves, L. M., Osborne, D., Robertson, A., Brunton, C., Armstrong, G., & Sibley, C. G. (2013). Sense of Community in New Zealand Neighbourhoods: A Multi-Level Model Predicting Social Capital. *New Zealand Journal of Psychology*, *42*(1), 36–45.

Shannon, H., Bush, K., Villeneuve, P., Hellemans, K., & Guimond, S. (2022). Problematic social media use in adolescents and young adults: A meta-analysis. *JMIR Mental Health*, *9*(4). https://doi.org/10.2196/33450

Shensa, A., Escobar-Viera, C., Sidani, J., Bowman, N., Marshal, M., & Primack, B. (2017). *Problematic Social Media Use and Depressive Symptoms Among U.S. Young Adults: A Nationally-Representative Study*. Social Science & Medicine. https://pubmed.ncbi.nlm.nih.gov/28446367/

Sibley, C. G. (2021). *Sampling procedure and sample details for the New Zealand Attitudes and Values Study*. https://doi.org/10.31234/osf.io/wgqvy

Sibley, C. G., Luyten, N., Purnomo, M., Mobberley, A., Wootton, L. W., Hammond, M. D., Sengupta, N., Perry, R., West-Newman, T., Wilson, M. S., McLellan, L., Hoverd, W. J., & Robertson, A. (2011). The Mini-IPIP6: Validation and extension of a short measure of the Big-Six factors of personality in New Zealand. *New Zealand Journal of Psychology*, *40*(3), 142–159.

Silva, R. C., & Steins, G. (2023). Social media and body dissatisfaction in young adults: An experimental investigation of the effects of different image content and influencing constructs. *Frontiers in Psychology*, *14*(1037932). https://doi.org/10.3389/fpsyg.2023.1037932

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, *53*(1), 80–93. https://doi.org/10.1037/0022-0167.53.1.80

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., & Wager, S. (2024). *Policytree: Policy learning via doubly robust empirical welfare maximization over trees*. https://CRAN.R-project.org/package=policytree

Tibshirani, J., Athey, S., Sverdrup, E., & Wager, S. (2024). *Grf: Generalized random forests*. https://github.com/grf-labs/grf

Twenge, J. M., & Campbell, W. K. (2018). Associations between Screen Time and Lower Psychological well-being among Children and adolescents: Evidence from a population-based Study. *Preventive Medicine Reports*, *12*, 271–283. https://doi.org/10.1016/j.pmedr.2018.10.003

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2017). Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates among U.S. Adolescents after 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science*, *6*(1), 3–17. https://doi.org/10.1177/2167702617723376

Valkenburg, P. M. (2022). Social Media Use and well-being: What We Know and What We Need to Know. *Current Opinion in Psychology*, *45*. https://doi.org/10.1016/j.copsyc.2021.12.006

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, *167*(4), 268–274. https://doi.org/10.7326/M16-2607

VanderWeele, T. J., Mathur, M. B., & Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statistical Science*, *35*(3), 437–466.

Verbrugge, L. M. (1997). A global disability indicator. *Journal of Aging Studies*, *11*(4), 337–362. https://doi.org/10.1016/S0890-4065(97)90026-8

Verbrugge, L. M., Rennert, C., & Madans, J. H. (1997). The great efficacy of personal and equipment assistance in reducing disability. *American Journal of Public Health*, *87*(3), 384–392. https://doi.org/10.2105/ajph.87.3.384

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.

Whitehead, J., Davie, G., Graaf, B. de, Crengle, S., Lawrenson, R., Miller, R., & Nixon, G. (2023). Unmasking hidden disparities: A comparative observational study examining the impact of different rurality classifications for health research in Aotearoa New Zealand. *BMJ Open*, *13*(4), e067927.

World Health Organization. (2024). *Teens, screens and mental health*. Who.int; World Health Organization: WHO. https://www.who.int/europe/news/item/25-09-2024-teens--screens-and-mental-health

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., & Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv Preprint arXiv:2111.07966*. https://doi.org/10.48550/arXiv.2111.07966

Yang, W., Morita, N., Zuo, Z., Kawaida, K., Ogai, Y., Saito, T., & Hu, W. (2021). Maladaptive Perfectionism and Internet Addiction among Chinese College Students: A Moderated Mediation Model of Depression and Gender. *International Journal of Environmental Research and Public Health*, *18*(5), 2748. https://doi.org/10.3390/ijerph18052748

Appendix A: Measures

Baseline Covariate Measures

Age
What is your date of birth?
We asked participants' ages in an open-ended question ("What is your age?" or "What is your date of birth"). (**string_is?** Developed for the NZAVS.)

Agreeableness
I sympathise with others' feelings. I am not interested in other people's problems. I feel others' emotions. I am not really interested in others (reversed).
Mini-IPIP6 Agreeableness dimension: (i) I sympathise with others' feelings. (ii) I am not interested in other people's problems. (r) (iii) I feel others' emotions. (iv) I am not really interested in others. (r) (Sibley et al., 2011)

Alcohol Frequency
"How often do you have a drink containing alcohol?"
Participants could chose between the following responses: '(1 = Never - I don't drink, 2 = Monthly or less, 3 = Up to 4 times a month, 4 = Up to 3 times a week, 5 = 4 or more times a week, 6 = Don't know)' (Health, 2013)

Alcohol Intensity
"How many drinks containing alcohol do you have on a typical day when drinking alcohol? (number of drinks on a typical day when drinking)"
Participants responded using an open-ended box. (Health, 2013)

Social Belonging
Know that people in my life accept and value me. Feel like an outsider (reversed). Know that people around me share my attitudes and beliefs.
We assessed felt belongingness with three items adapted from the Sense of Belonging Instrument (Hagerty & Patusky, 1995): (1) "Know that people in my life accept and value me"; (2) "Feel like an outsider"; (3) "Know that people around me share my attitudes and beliefs". Participants responded on a scale from 1 (Very Inaccurate) to 7 (Very Accurate). The second item was reversely coded. (Hagerty & Patusky, 1995)

Born in NZ
Where were you born? (please be specific, e.g., which town/city?)
Coded binary (1 = New Zealand; 0 = elsewhere.) (**string_is?** Developed for the NZAVS.)

Conscientiousness
I get chores done right away. I like order. I make a mess of things. I often forget to put things back in their proper place.

Mini-IPIP6 Conscientiousness dimension: (i) I get chores done right away. (ii) I like order. (iii) I make a mess of things. (r) (iv) I often forget to put things back in their proper place. (r) (Sibley et al., 2011)

Education Level
What is your highest level of qualification?
We asked participants, "What is your highest level of qualification?". We coded participans highest finished degree according to the New Zealand Qualifications Authority. Ordinal-Rank 0-10 NZREG codes (with overseas school qualifications coded as Level 3, and all other ancillary categories coded as missing) (**string_is?** Developed for the NZAVS.)

Employed
Are you currently employed (This includes self-employed of casual work)?
Binary response: (0 = No, 1 = Yes) (**string_is?** Stats NZ Census Question)

Ethnicity
Which ethnic group(s) do you belong to?
Coded string: (1 = New Zealand European; 2 = Māori; 3 = Pacific; 4 = Asian) (**string_is?** NZ Census coding.)

Disability Status
Do you have a health condition or disability that limits you and that has lasted for 6+ months?
We assessed disability with a one-item indicator adapted from Verbrugge (1997). It asks, "Do you have a health condition or disability that limits you and that has lasted for 6+ months?" (1 = Yes, 0 = No). (Verbrugge, 1997)

Log Hours with Children
Hours spent…looking after children.
We took the natural log of the response + 1. (Sibley et al., 2011)

Log Hours Commuting
Hours spent…travelling/commuting.
We took the natural log of the response + 1. (**string_is?** Developed for the NZAVS.)

Log Hours of Exercise
Hours spent…exercising/physical activity.
We took the natural log of the response + 1. (Sibley et al., 2011)

Log Hours on Housework
Hours spent…housework/cooking.
We took the natural log of the response + 1. (Sibley et al., 2011)

Log Household Income
Please estimate your total household income (before tax) for the year XXXX.
We took the natural log of the response + 1. (**string_is?** Developed for the NZAVS.)

Male
We asked participants' gender in an open-ended question: "what is your gender?"
Here, we coded all those who responded as Male as 1, and those who did not as 0. (Fraser et al., 2020)

Neuroticism
I have frequent mood swings. I am relaxed most of the time (reversed). I get upset easily. I seldom feel blue (reversed).
Mini-IPIP6 Neuroticism dimension: (i) I have frequent mood swings. (ii) I am relaxed most of the time. (r) (iii) I get upset easily. (iv) I seldom feel blue. (r) (Sibley et al., 2011)

Non-Heterosexual
How would you describe your sexual orientation? (e.g., heterosexual, homosexual, straight, gay, lesbian, bisexual, etc.)
Open-ended question coded as binary (not heterosexual = 1). (Greaves et al., 2017)

NZ Deprivation Index
New Zealand Deprivation - Decile Index - Using 2018 Census Data
Numerical: (1-10) (Atkinson et al., 2019)

Occupational Prestige Index
We assessed occupational prestige and status using the New Zealand Socio-economic Index 13 (NZSEI-13).
This index uses the income, age, and education of a reference group, in this case, the 2013 New Zealand census, to calculate a score for each occupational group. Scores range from 10 (Lowest) to 90 (Highest). This list of index scores for occupational groups was used to assign each participant a NZSEI-13 score based on their occupation. (Fahy et al., 2017)

Openness
I have a vivid imagination. I have difficulty understanding abstract ideas (reversed). I do not have a good imagination (reversed). I am not interested in abstract ideas (reversed).
Mini-IPIP6 Openness to Experience dimension: (i) I have a vivid imagination. (ii) I have difficulty understanding abstract ideas. (r) (iii) I do not have a good imagination. (r) (iv) I am not interested in abstract ideas. (r) (Sibley et al., 2011)

Parent
If you are a parent, in which year was your eldest child born?
Parents were coded as 1, while the others were coded as 0. (**Developed?** for the NZAVS.)

Has Partner
What is your relationship status? (e.g., single, married, de-facto, civil union, widowed, living together, etc.)
Coded as binary (has partner = 1). (**string_is?** Developed for the NZAVS.)

Political Conservatism
Please rate how politically liberal versus conservative you see yourself as being.
Ordinal response: (1 = Extremely Liberal, 7 = Extremely Conservative) (Jost, 2006)

Major Religions
Do you identify with a religion and/or spiritual group? –> (If yes…)–> What religion or spiritual group?
Open-ended (string). Coded from New Zealand Census Categories. Levels are: "Not Religious","Anglican","Buddhist", "Catholic", "Christian (Non-Denominational)", "Christian (Other Denominations)","Hindu", "Jewish", "Muslim","Presbyterian, Congregational, Reformed", "Other Religions". (**coded?** for the NZAVS.)

Religious Identification
How important is your religion to how you see yourself?
Ordinal response: (1 = Not Important, 7 = Very Important) (**string_is?** Developed for the NZAVS.)

Rural Classification
High Urban Accessibility = 1, Medium Urban Accessibility = 2, Low Urban Accessibility = 3, Remote = 4, Very Remote = 5.
"Participants residence locations were coded according to a five-level ordinal categorisation ranging from Urban to Rural." (Whitehead et al., 2023)

Sample Frame Opt in
Participant was not randomly sampled from the New Zealand Electoral Roll.
Code string (Binary): (0 = No, 1 = Yes) (**string_is?** Developed for the NZAVS.)

Short Form Health
In general, would you say your health is…
Ordinal response: (1 = Poor, 7 = Excellent) (Instrument Ware Jr & Sherbourne, 1992)

Smoker
Do you currently smoke tobacco cigarettes?
Binary smoking indicator (0 = No, 1 = Yes). (**string_is?** Developed for NZAVS.)

Exposure Measure

Log Hours on Social Media
Hours spent … using social media (e.g., Facebook)

We took the natural log of the response + 1. (Sibley et al., 2011)

Outcome Measures

Anxiety
During the past 30 days, how often did…you feel restless or fidgety? During the past 30 days, how often did…you feel that everything was an effort? During the past 30 days, how often did…you feel nervous?
Ordinal response: (0 = None Of The Time; 1 = A Little Of The Time; 2= Some Of The Time; 3 = Most Of The Time; 4 = All Of The Time) (Kessler et al., 2002)

Depression
During the past 30 days, how often did…you feel hopeless? During the past 30 days, how often did…you feel so depressed that nothing could cheer you up? During the past 30 days, how often did…you feel you feel restless or fidgety?
Ordinal response: (0 = None Of The Time; 1 = A Little Of The Time; 2= Some Of The Time; 3 = Most Of The Time; 4 = All Of The Time) (Kessler et al., 2002)

Life Satisfaction
I am satisfied with my life. In most ways my life is close to ideal.
Ordinal response (1 = Strongly Disagree to 7 = Strongly Agree). (Diener et al., 1985)

Personal Well-Being Index
How satisfied are you with your standard of living? How satisfied are you with your health? How satisfied are you with what you are achieving in life? How satisfied are you with your personal relationships? How satisfied are you with how safe you feel? How satisfied are you with feeling part of your community? How satisfied are you with your future security?
Ordinal response: (0 = completely dissatisfied to 10 = completely satisfied). (Cummins et al., 2003)

Rumination
During the last 30 days, how often did…you have negative thoughts that repeated over and over?
Ordinal responses: 0 = None of The Time, 1 = A little of The Time, 2 = Some of The Time, 3 = Most of The Time, 4 = All of The Time. (Nolen-hoeksema & Morrow, 1993)

Self Esteem
On the whole am satisfied with myself. Take a positive attitude toward myself. Am inclined to feel that I am a failure (reversed).
Ordinal response (1 = Very inaccurate to 7 = Very accurate). (Rosenberg, 1965)

Body Satisfaction
Am satisfied with the appearance, size and shape of my body.
Ordinal response (1 = Very inaccurate to 7 = Very accurate). (Rosenberg, 1965)

Perfectionism

Doing my best never seems to be enough. My performance rarely measures up to my standards. I am hardly ever satisfied with my performance.

Ordinal response (1 = Very inaccurate to 7 = Very accurate). (Rice, Richardson, & Tueller, 2014)

Meaning in Life

My life has a clear sense of purpose. I have a good sense of what makes my life meaningful.

Ordinal response (1 = Very inaccurate to 7 = Very accurate). (Steger, Frazier, Oishi, & Kaler, 2006)

Appendix B: Sample Characteristics

Baseline Confounders

Variable	2018
	(N=23594)
Age	
Mean (SD)	44.1 (13.3)
Median [Min, Max]	45.0 [18.0, 65.0]
Agreeableness	
Mean (SD)	5.42 (0.969)
Median [Min, Max]	5.50 [1.00, 7.00]
Missing	202 (0.9%)
Alcohol Frequency	
Mean (SD)	2.11 (1.30)
Median [Min, Max]	2.00 [0, 5.00]
Missing	173 (0.7%)
Alcohol Intensity	
Mean (SD)	2.29 (2.19)
Median [Min, Max]	2.00 [0, 15.0]
Missing	758 (3.2%)
Belong	
Mean (SD)	5.12 (1.07)
Median [Min, Max]	5.31 [1.00, 7.00]
Missing	200 (0.8%)
Born in New Zealand	
0	5042 (21.4%)
1	18263 (77.4%)
Missing	289 (1.2%)
Conscientiousness	
Mean (SD)	5.07 (1.07)
Median [Min, Max]	5.21 [1.00, 7.00]
Missing	199 (0.8%)
Education Level	
no_qualification	447 (1.9%)
cert_1_to_4	7918 (33.6%)
cert_5_to_6	2724 (11.5%)
university	6728 (28.5%)
post_grad	2734 (11.6%)
masters	1990 (8.4%)
doctorate	480 (2.0%)
Missing	573 (2.4%)
Employed	

0	3902 (16.5%)
1	19669 (83.4%)
Missing	23 (0.1%)
Ethnicity	
euro	18496 (78.4%)
maori	2860 (12.1%)
pacific	599 (2.5%)
asian	1406 (6.0%)
Missing	233 (1.0%)
Disability Status	
Mean (SD)	0.203 (0.402)
Median [Min, Max]	0 [0, 1.00]
Log Hours with Children	
Mean (SD)	1.34 (1.71)
Median [Min, Max]	0.0387 [0, 5.13]
Log Hours Commuting	
Mean (SD)	1.55 (0.808)
Median [Min, Max]	1.61 [0, 4.40]
Log Hours Exercising	
Mean (SD)	1.51 (0.818)
Median [Min, Max]	1.61 [0, 4.40]
Log Hours on Housework	
Mean (SD)	2.18 (0.745)
Median [Min, Max]	2.30 [0, 5.13]
Log Household Income	
Mean (SD)	11.4 (0.732)
Median [Min, Max]	11.5 [0.704, 14.4]
Missing	1829 (7.8%)
Male	
0	16462 (69.8%)
1	7057 (29.9%)
Missing	75 (0.3%)
Neuroticism	
Mean (SD)	3.60 (1.15)
Median [Min, Max]	3.53 [1.00, 7.00]
Missing	202 (0.9%)
Non-heterosexual	
0	20632 (87.4%)
1	1905 (8.1%)
Missing	1057 (4.5%)
NZ Deprivation Index	
Mean (SD)	4.79 (2.71)

Median [Min, Max]	4.95 [1.00, 10.0]
Missing	194 (0.8%)
Occupational Prestige Index	
Mean (SD)	54.5 (16.5)
Median [Min, Max]	56.0 [10.0, 90.0]
Missing	265 (1.1%)
Openness	
Mean (SD)	5.00 (1.11)
Median [Min, Max]	5.02 [1.00, 7.00]
Missing	200 (0.8%)
Parent	
0	7949 (33.7%)
1	15437 (65.4%)
Missing	208 (0.9%)
Has Partner	
Mean (SD)	0.743 (0.437)
Median [Min, Max]	1.00 [0, 1.00]
Missing	738 (3.1%)
Political Conservatism	
Mean (SD)	3.49 (1.37)
Median [Min, Max]	3.96 [1.00, 7.00]
Missing	1469 (6.2%)
Religious Identification	
Mean (SD)	2.28 (2.13)
Median [Min, Max]	1.00 [1.00, 7.00]
Missing	589 (2.5%)
Rural Classification	
High Urban Accessibility	14938 (63.3%)
Medium Urban Accessibility	4272 (18.1%)
Low Urban Accessibility	2720 (11.5%)
Remote	1219 (5.2%)
Very Remote	252 (1.1%)
Missing	193 (0.8%)
Sample Frame Opt-In	
0	22832 (96.8%)
1	762 (3.2%)
Short Form Health	
Mean (SD)	5.03 (1.18)
Median [Min, Max]	5.04 [1.00, 7.00]
Smoker	
0	21830 (92.5%)
1	1764 (7.5%)

Table 4: *Baseline confounder demographic statistics for New Zealand Attitudes and Values Cohort.*

Exposure Variable

Variable	2018	2020
Sample Size	N = 23,594	N = 23,594
Hours on Social Media		
Mean (SD)	6.66 (8.85)	6.29 (7.83)
Median [Min, Max]	4.02 [1.00, 150]	4.03 [0, 168]
Missing	0 (0%)	8,800 (37.3%)
Hours on Social Media (binary)		
[0.0, 14.0)	21,122 (89.5%)	13,272 (56.3%)
[14.0, 168.0]	2,472 (10.5%)	1,522 (6.5%)
Missing	0 (0%)	8,800 (37.3%)

Table 5: *Exposure variable demographic statistics for New Zealand Attitudes and Values Cohort.*

Outcome Variables

Variable	2018	2022	Overall
Sample Size	N = 23,594	23,594	N = 47,188
Body Satisfaction			
Mean (SD)	4.09 (1.72)	4.04 (1.71)	4.07 (1.71)
Median [Min, Max]	4.02 [1.00, 7.00]	4.01 [1.00, 7.00]	4.02 [1.00, 7.00]
Missing	260 (1.1%)	12,941 (54.8%)	13,201 (28.0%)
Anxiety			
Mean (SD)	1.30 (0.776)	1.26 (0.773)	1.29 (0.775)
Median [Min, Max]	1.33 [0, 4.00]	1.33 [0, 4.00]	1.33 [0, 4.00]
Missing	215 (0.9%)	12,421 (52.6%)	12,636 (26.8%)
Depression			
Mean (SD)	0.640 (0.771)	0.588 (0.742)	0.623 (0.762)
Median [Min, Max]	0.333 [0, 4.00]	0.333 [0, 4.00]	0.333 [0, 4.00]
Missing	217 (0.9%)	12,421 (52.6%)	12,638 (26.8%)
Life Meaning			
Mean (SD)	5.41 (1.18)	5.41 (1.21)	5.41 (1.19)
Median [Min, Max]	5.52 [1.00, 7.00]	5.53 [1.00, 7.00]	5.52 [1.00, 7.00]
Missing	3 (0.0%)	12,545 (53.2%)	12,548 (26.6%)
Life Satisfaction			
Mean (SD)	5.27 (1.21)	5.15 (1.23)	5.23 (1.22)
Median [Min, Max]	5.50 [1.00, 7.00]	5.47 [1.00, 7.00]	5.49 [1.00, 7.00]
Missing	121 (0.5%)	12,707 (53.9%)	12,828 (27.2%)
Perfectionism			

Mean (SD)	3.26 (1.36)	3.09 (1.43)	3.20 (1.38)
Median [Min, Max]	3.04 [1.00, 7.00]	2.98 [1.00, 7.00]	3.02 [1.00, 7.00]
Missing	2 (0.0%)	12,473 (52.9%)	12,475 (26.4%)
Personal Well-being Index			
Mean (SD)	7.04 (1.64)	6.88 (1.71)	6.99 (1.66)
Median [Min, Max]	7.26 [0, 10.0]	7.21 [0, 10.0]	7.25 [0, 10.0]
Missing	23 (0.1%)	12,399 (52.6%)	12,422 (26.3%)
Rumination			
Mean (SD)	0.922 (1.03)	0.841 (0.976)	0.896 (1.01)
Median [Min, Max]	0.967 [0, 4.00]	0.956 [0, 4.00]	0.963 [0, 4.00]
Missing	254 (1.1%)	12,448 (52.8%)	12,702 (26.9%)
Self Esteem			
Mean (SD)	5.05 (1.30)	5.13 (1.32)	5.07 (1.31)
Median [Min, Max]	5.30 [1.00, 7.00]	5.35 [1.00, 7.00]	5.32 [1.00, 7.00]
Missing	202 (0.9%)	12,438 (52.7%)	12,640 (26.8%)

Table 6: *Outcome variables measured at 2018 and 2022.*

Appendix C: Transition Matrix to Check the Positivity Assumption

From / To	State 0	State 1	Total
State 0	12731	905	13636
State 1	695	463	1158

***Table 7**: Transition Matrix Showing Change.*

These transition matrices capture shifts in states between consecutive waves. Each cell shows the count of individuals transitioning from one state to another. Rows are the initial state (From) and columns are the subsequent state (To).

Appendix D: RATE AUTOC and RATE Qini

RATE Test

The RATE metric shows how much extra gain (or avoided loss) we achieve by **targeting** instead of treating everyone identically.

Technical note: In code we always set `policy = "treat_best"`; for harmful exposures this is interpreted as *'treat-those-most-sensitive'* (i.e., prioritise protection or withholding).

- **Beneficial exposure:** we rank by positive CATEs and deliver the exposure to those predicted to **benefit most**.
- **Detrimental exposure:** we rank by increasingly **positive** CATEs (more predicted harm) and identify those who should be protected or withheld from the exposure.

Either way, a larger **absolute** RATE shows that a CATE-based targeting rule 'outperforms' a one-size-fits-all policy—by boosting outcomes for beneficial exposures or – in the case where we are explore sensitivity to harm – evaluating increasing harms for detrimental ones.

Recall we flipped Anxiety, Depression, Rumination so **'higher' always tracks the analysis goal: higher = more benefit for beneficial exposures, higher = more harm for detrimental exposures.**

Because we test several outcomes, RATE *p*-values are adjusted with Benjamini–Hochberg false-discovery-rate adjustment (q = 0.1) before we decide whether heterogeneity is actionable.

Comparison of targeting operating characteristic (TOC) by rank average treatment effect (RATE): AUTOC vs QINI

We applied two TOC by RATE methods to the same causal-forest $\tau(x)$ estimates:

- **AUTOC** intensifies focus on top responders via logarithmic weighting.
- **QINI** balances effect size and prevalence via linear weighting.

Exploratory RATE analysis; controlled FDR at q=0.20 over 12 outcomes.
Both methods yield positive RATE estimates for: **Hours of Exercise (log)**.
This concordance indicates robust heterogeneity evidence.

When methods disagree (only QINI yields positive RATE for Anxiety (reversed)), choose **QINI** for overall benefit or **AUTOC** to focus top responders.

RATE AUTOC Results

Evidence for heterogeneous treatment effects (policy = treat best responders) using AUTOC

AUTOC uses logarithmic weighting to focus treatment on top responders.

Note: The following outcomes were inverted during preprocessing because higher values of the exposure correspond to worse outcomes: Body Satisfaction, Self-Esteem, Life Satisfaction, Personal Well-being Index, Life Meaning.

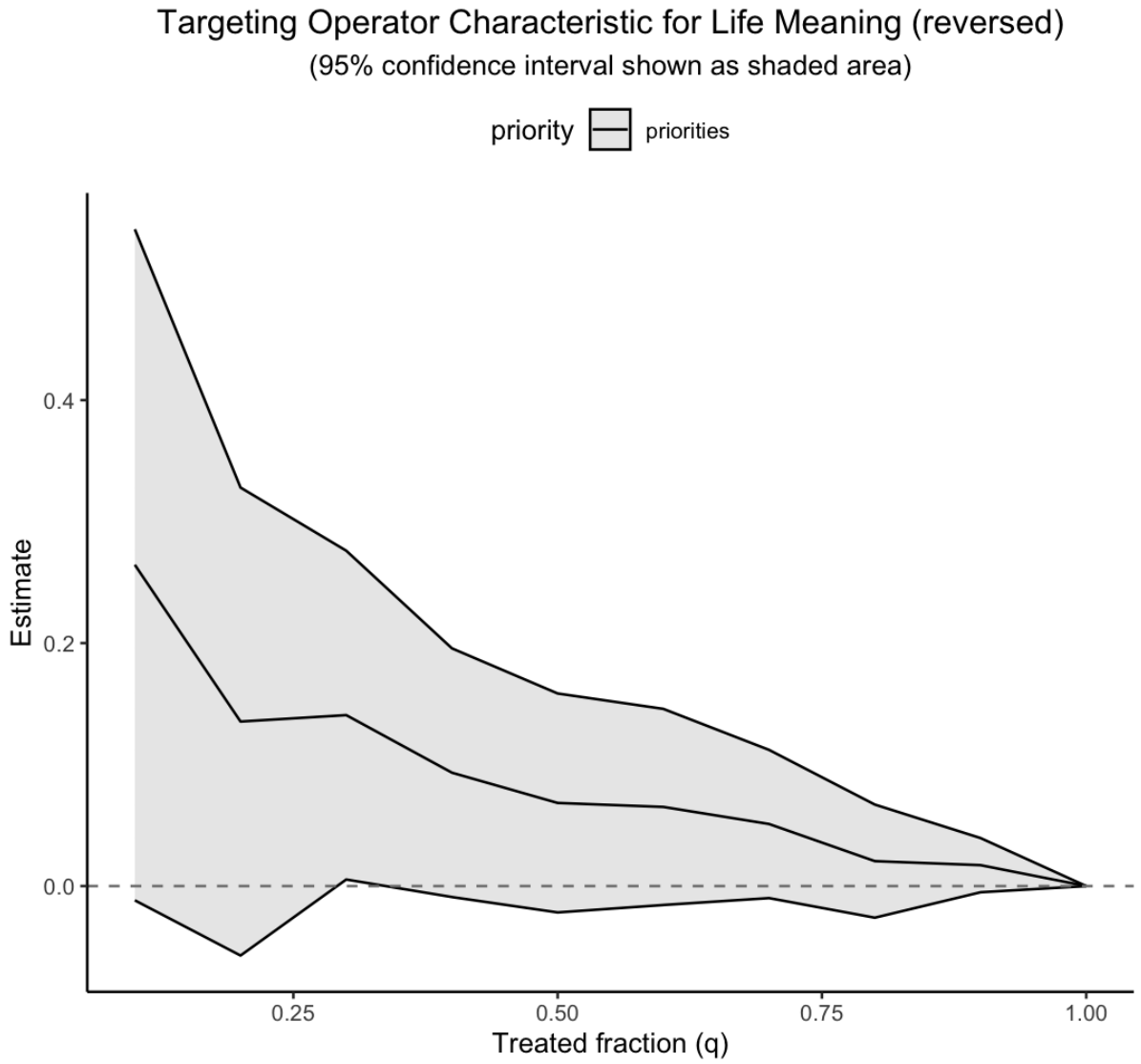
Positive RATE estimates for: **Life Meaning (reversed)**.

Estimates (**Life Meaning (reversed)**: 0.080 (95% CI -0.000, 0.160)) show robust heterogeneity.

Negative RATE estimates for: Anxiety, Rumination, Depression.

Estimates (Anxiety: -0.071 (95% CI -0.144, 0.002); Rumination: -0.081 (95% CI -0.171, 0.009); Depression: -0.166 (95% CI -0.307, -0.025)) caution against CATE prioritisation.

For outcomes with adjusted p-values not meeting the FDR threshold of q = 0.20 (Personal Well-being Index (reversed), Life Satisfaction (reversed), Self Esteem (reversed), Body Satisfaction (reversed), Perfectionism), evidence is inconclusive.



Appendix E: Approach to Heterogeneous Treatment Effects

Appendix X. Estimating and Interpreting Heterogeneous Treatment Effects with **grf**

Here we explain a heterogeneous-treatment-effect (HTE) analysis using causal forests (Tibshirani et al., 2024). In our workflow, we move from the average treatment effect (ATE) to individualised effects, quantify the practical value of targeting, and finish with interpretable decision rules.

1 Average Treatment Effect (ATE)

The ATE answers: *'What would happen, on average, if everyone received treatment versus no one?'*

$$\text{ATE} = E[Y(1) - Y(0)].$$

Using the grf package, we estimate the ATE doubly-robustly. Because we analyse several outcomes, we adjust ATE *p*-values with bonferroni (α = 0.05) to control the family-wise error rate.

2 Do Effects Vary? Formal Test of Heterogeneity

Define the conditional average treatment effect (CATE)

$$\tau(x) = E[Y(1) - Y(0) \mid X = x].$$

If $\tau(x)$ is constant, effects are homogeneous; otherwise they vary. Classical interaction models impose strong forms; **grf** uses *causal forests* to discover complex, nonlinear heterogeneity (Wager & Athey, 2018). We assess heterogeneity with RATE *p*-values corrected via Benjamini–Hochberg false-discovery-rate adjustment (q = 0.1), controlling the false-discovery rate (Benjamini & Hochberg, 1995).

3 Causal Forests for Individualised Estimates

A causal forest is an ensemble of 'honest' causal trees that split on covariates to maximise treated–control contrasts. For each unit i we obtain

$$\hat{\tau}(x_i)$$

Strengths are flexibility, orthogonalisation, and per-person estimates.

4 Built-in Protection Against Over-fitting

Honesty (split half/estimate half) plus out-of-bag (OOB) predictions yield unbiased $\hat{\tau}(x)$ and standard errors without manual hyper-tuning.

5 Missing Data Handling

grf deploys 'Missing Incorporated in Attributes' (MIA): missingness is a valid split, so cases stay in the analysis – no ad-hoc imputation required.

6 Testing for **Actionable** Heterogeneity: the TOC & RATE Metrics

Ranking units by $\hat{\tau}$ defines a **Targeting Operator Characteristic** (TOC) curve: the cumulative gain from treating the top fraction q of predicted responders. Two scalar summaries:

- **RATE AUTOC** – area under the entire TOC; emphasises the very highest responders.
- **RATE Qini** – weighted area with weight q; rewards sustained gains across larger coverage (Yadlowsky et al., 2021).

Under H_0: $\tau(x)$ constant, both equal 0. grf::rank_average_treatment_effect() supplies point estimates, standard errors, and t-tests.

Multiplicity control: We adjust AUTOC and Qini p-values with Benjamini–Hochberg false-discovery-rate adjustment (q = 0.1) before declaring actionable heterogeneity.
Here is an **interpretation tip**:

- AUTOC answers *'How sharply can we prioritise?'*
- Qini answers *'How valuable is targeting when budgets are modest but not tiny?'*

7 Visualising Policy Value: the Qini Curve
Plotting the Qini curve (cumulative gain vs q) reveals where returns plateau. Investigators (and policy audiences) can see at a glance whether benefits concentrate in, say, the top 20 % or persist up to 50 %.
8 Valid Inference for RATE / Qini
Although OOB predictions are out-of-sample per tree, they inherit forest-level dependence. We use an explicit **sample split**:

1. **Train set**: fit the causal forest and compute $\hat{\tau}(x)$.
2. **Test set**: compute RATE AUTOC/Qini and run H_0 tests.

This second split yields honest policy evaluation and guards against optimistic bias (Tibshirani et al., 2024).
9 From Black Box to Simple Rules: Policy Trees
Stakeholders value transparent criteria. The **policytree** algorithm takes $\hat{\tau}(x)$ or doubly-robust scores and learns a shallow decision tree that maximises expected welfare (Sverdrup et al., 2024).
Advantages: interpretability, the possibility of fairness constraints, and easy communication (e.g., *'treat if age < 25 and baseline severity high'*).
Training mirrors the split above: learn the tree on one fold, evaluate welfare on another.

> **Caveat** Splits identify predictors of *effect variation*, not causal levers. Changing a covariate in the tree does **not** guarantee an effect on $\tau(x)$.

10 Ethical and Practical Considerations
Statistical optimisation rarely aligns perfectly with equity or political feasibility. Decisions about who *should* receive treatment belong to democratic processes that weigh fairness, cost, and broader social values.
Putting it together
The sequence—ATE, causal-forest CATEs, RATE/Qini diagnostics, Qini curve, and finally a shallow policy tree—delivers both rigorous evidence and a defensible targeting rule. Researchers learn **how large** heterogeneity is, **where** targeting pays off under budget constraints, and **which** simple covariate splits capture most of the welfare gain, all while guarding against over-fitting and multiplicity.

Appendix F: Strengths and Limitations of Causal Forests

Strengths and Limitations of Our Approach

We used causal forests (Tibshirani et al., 2024) to estimate how treatment effects may differ for individuals with different characteristics. This method is powerful, however it also depends on measuring all major variables influencing both treatment selection and outcomes. If such variables are missed or mismeasured, results can be biased. Additionally, interpreting subgroup effects can be tricky when many characteristics are involved and statistically significant differences may not always translate into meaningful real-world gains.

Despite these concerns, causal forests offer notable advantages. They allow for flexible, non-parametric modelling (Tibshirani et al., 2024), avoiding strict assumptions that might miss complex interactions. We used a robust evaluation method—training our model on half the data and testing it on the remaining half—to avoid overfitting. We then checked whether the predicted differences were genuine and estimated how much benefit we might gain by targeting treatment to those likely to benefit most (Tibshirani et al., 2024; Wager & Athey, 2018). Qini curves (Tibshirani et al., 2024) let us see the overall improvement from treating the top-ranked individuals first, and policy trees (Athey & Wager, 2021b, 2021a; Sverdrup et al., 2024) turn these findings into simple 'if-then' rules. Together, this approach provides a practical means of identifying and acting on genuine treatment effect differences.

Appendix G: R Code

Script 1:

```
# script 1 workflow lecture 10
# may 2025
# questions: joseph.bulbulia@vuw.ac.nz


# +-------------------------+
# |      DO NOT ALTER       |
# +-------------------------+

# restart fresh session for a clean workspace
rstudioapi::restartSession()

# set seed for reproducibility
set.seed(123)

# essential library ------------------------------------------------------
# install and load 'margot' from GitHub if missing
if (!require(margot, quietly = TRUE)) {
  devtools::install_github("go-bayes/margot")
  library(margot)
}


if (packageVersion("margot") < "1.0.44") {
  stop("please install margot >= 1.0.44 for this workflow\n
       run: devtools::install_github(\"go-bayes/margot\")
")
}

# call library
library("margot")

# load packages ----------------------------------------------------------
# pacman will install missing packages automatically
if (!requireNamespace("pacman", quietly = TRUE)) install.packages("pacman")
pacman::p_load(
  tidyverse,       # data wrangling + plotting
  qs,              # fast data i/o
  here,            # project-relative file paths
  data.table,      # fast data manipulation
  fastDummies,     # dummy variable creation
  naniar,          # missing data handling
  skimr,           # summary statistics
  grf,             # machine learning forests
  kableExtra,      # tables
  ggplot2,         # graphs
```

```
  doParallel,       # parallel processing
  grf,             # causal forests
  janitor,          # variables names
  stringr,          # variable names
  patchwork,        # graphs
  table1,          # tables,
  cli
)


# create directories ------------------------------------------------------
# create data directory if it doesn't exist
if (!dir.exists("data")) {
  dir.create("data")  # first time only: make a folder named 'data'
}

if (!dir.exists("save_directory")) {
  dir.create("save_directory")  # first time only: make a folder named 'data'
}

# set up data directory structure
data_dir    <- here::here("data")
push_mods <- here::here("save_directory")


# load data ---------------------------------------------------------------
df_nz_long <- margot::here_read_qs("df_nz_long", data_dir)

df_nz_long$hours_social_media

#check variable is there
table1::table1(~ hours_social_media | wave, data = df_nz_long)

# initial data prep -------------------------------------------------------
# prepare intial data
# define labels for rural classification
rural_labels <- c(
  "High Urban Accessibility",
  "Medium Urban Accessibility",
  "Low Urban Accessibility",
  "Remote",
  "Very Remote"
)

dat_prep <- df_nz_long |>
  arrange(id, wave) |>
  margot::remove_numeric_attributes() |>
  mutate(
    # cap extreme values
    alcohol_intensity = pmin(alcohol_intensity, 15),
```

```
    # flag heavy drinkers: freq ≥3 → 1, ≤2 → 0, else NA
    heavy_drinker = case_when(
      alcohol_frequency >= 3 ~ 1,
      alcohol_frequency <= 2 ~ 0,
      TRUE                   ~ NA_real_
    ),
    # map freq categories to weekly counts
    alcohol_frequency_weekly = recode(
      alcohol_frequency,
      `0` = 0, `1` = 0.25,
      `2` = 1, `3` = 2.5,
      `4` = 4.5,
      .default = NA_real_
    ),
    # relabel rural factor
    rural_gch_2018_l = factor(
      rural_gch_2018_l,
      levels = 1:5,
      labels = rural_labels,
      ordered = TRUE
    )
  ) |>
  droplevels()



# view variable names ------------------------------------------------------
print(colnames(df_nz_long))


# get total participants
n_total = length(unique(df_nz_long$id))



# pretty number
n_total = margot::pretty_number(n_total)

# save
here_save(n_total, "n_total")

# +--------------------------+
# |     END DO NOT ALTER     |
# +--------------------------+



# +--------------------------+
# |    MODIFY THIS SECTION   |
# +--------------------------+
```

```
# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# | OPTIONALLY MODIFY SECTION|
# +-------------------------+

# define study variables ---------------------------------------------------
# ** key decision 1: define your three study waves **
# **  define your study waves **
baseline_wave      <- "2018"        # baseline measurement
exposure_waves     <- c("2020")     # when exposure is measured
outcome_wave       <- "2022"        # when outcomes are measured
all_waves          <- c(baseline_wave, exposure_waves, outcome_wave)

cli::cli_h1("set waves for three-wave study ✓")


# +-------------------------+
# |END OPTIONALLY MODIFY SEC.|
# +-------------------------+
# +-------------------------+
# |       END ALERT         |
# +-------------------------+



# define exposure variable -------------------------------------------------
# ** key decision 2: define your exposure variable **

# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# |    MODIFY THIS SECTION   |
# +-------------------------+
name_exposure <- "hours_social_media"

# exposure variable labels
var_labels_exposure <- list(
  "hours_social_media" = "Hours on Social Media",
  "hours_social_media_binary" = "Hours on Social Media (binary)"
)

cli::cli_h1("set variable name for exposure ✓")

# +-------------------------+
# |       END ALERT         |
# +-------------------------+
```

```
# +-------------------------+
# |   END MODIFY SECTION    |
# +-------------------------+




# define outcome variables ----------------------------------------
# ** key decision 3: define your outcome variable **
# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# |    MODIFY THIS SECTION  |
# +-------------------------+
# ** key decision 3: define outcome variables **
# here, we are focussing on a subset of wellbeing outcomes
# chose outcomes relevant to * your * study. Might be all/some/none/exactly
# these:

outcome_vars <- c(
  "kessler_latent_anxiety",
  "kessler_latent_depression",
  "rumination",
  "bodysat",
  "lifesat",
  "perfectionism",
  "pwi",
  "self_esteem",
  "lifemeaning"
)

cli::cli_h1("set variable name for outcomes ✓")

# +-------------------------+
# |   END MODIFY SECTION    |
# +-------------------------+
# +-------------------------+
# |       END ALERT         |
# +-------------------------+


# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# | OPTIONALLY MODIFY SECTION|
# +-------------------------+
```

```
# define baseline variables ----------------------------------------------
# key decision 4 **  define baseline covariates **
# these are demographics, traits, etc. measured at baseline, that are common
# causes of the exposure and outcome.
# note we will automatically include baseline measures of the exposure and outcome
# later in the workflow.


baseline_vars <- c(
  # demographics
  "age", "born_nz_binary", "education_level_coarsen",
  "employed_binary", "eth_cat", "male_binary",
  "not_heterosexual_binary", "parent_binary", "partner_binary",
  "rural_gch_2018_l", "sample_frame_opt_in_binary",

  # personality traits (excluding exposure)
  "agreeableness", "conscientiousness", "neuroticism", "openness",

  # health and lifestyle
  "alcohol_frequency", "alcohol_intensity", "hlth_disability_binary",
  "log_hours_children", "log_hours_commute", "log_hours_exercise",
  "log_hours_housework", "log_household_inc",
  "short_form_health", "smoker_binary",

  # social and psychological
  "belong", "nz_dep2018", "nzsei_13_l",
  "political_conservative", "religion_identification_level"
)


cli::cli_h1("set baseline covariate names ✓")

# +--------------------------+
# |       END ALERT          |
# +--------------------------+
# +--------------------------+
# |   END MODIFY SECTION     |
# +--------------------------+




# +--------------------------+
# |      DO NOT ALTER        |
# +--------------------------+

# after selecting your exposure/ baseline / outcome variables do not modify this
# code
```

```
# make binary variable (UNLESS YOUR EXPOSURE IS A BINARY VARIABLE)
exposure_var_binary = paste0(name_exposure, "_binary")

# make exposure variable list (we will keep both the continuous and binary variable)
exposure_var  <- c(name_exposure, paste0(name_exposure, "_binary"))

# sort for easier reference
baseline_vars <- sort(baseline_vars)
outcome_vars <- sort(outcome_vars)

# save key variables -------------------------------------------------------
margot::here_save(name_exposure, "name_exposure")
margot::here_save(var_labels_exposure,"var_labels_exposure")
margot::here_save(baseline_vars,"baseline_vars")
margot::here_save(exposure_var, "exposure_var")
margot::here_save(exposure_var_binary, "exposure_var_binary")
margot::here_save(outcome_vars, "outcome_vars")
margot::here_save(baseline_wave, "baseline_wave")
margot::here_save(exposure_waves, "exposure_waves")
margot::here_save(outcome_wave, "outcome_wave")
margot::here_save(all_waves,"all_waves")

cli::cli_h1("saved names and labels to be used for manuscript ✓")


# +-------------------------+
# |    END DO NOT ALTER     |
# +-------------------------+



# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# | OPTIONALLY MODIFY SECTION|
# +-------------------------+


# Select eligible participants ---------------------------------------------
# only include participants who have exposure data at baseline

ids_baseline <- dat_prep |>
  # allow missing exposure at baseline
  # this would give us greater confidence that we generalise to the target population
  # filter(wave == baseline_wave) |>
  # option: do not allow missing exposure at baseline
  # this gives us greater confidence that we recover a incident effect
  filter(wave == baseline_wave, !is.na(!!sym(name_exposure))) |>
  pull(id)
```

```
ids_baseline <- dat_prep |>
  filter(wave == baseline_wave) |>
  # Must have social media hours data at baseline
  filter(!is.na(!!sym(name_exposure))) |>
  # Age restrictions appropriate for social media research
  filter(age >= 13 & age <= 65) |>
  # Focus on regular social media users (more than 1 hour per day)
  filter(!!sym(name_exposure) > 1) |>
  pull(id)

# filter data to include only eligible participants and relevant waves
dat_long_1 <- dat_prep |>
  filter(id %in% ids_baseline, wave %in% all_waves) |>
  droplevels()

# Count eligible participants
n_participants <- length(ids_baseline)

here_save(n_participants, "n_participants")
#n_participants_with_followup <- length(ids_baseline_final)
n_participants_pretty <- margot::pretty_number("n_participants")
# select eligible participants ----------------------------------------------
# only include participants who have exposure data at baseline

# You might require tighter conditions
# for example, if you are interested in the effects of hours of childcare,
# you might want to select only those who were parents at baseline.
# talk to me if you think you might night tighter eligibility criteria.

ids_baseline <- dat_prep |>
  # allow missing exposure at baseline
  # this would give us greater confidence that we generalise to the target population
  # filter(wave == baseline_wave) |>
  # option: do not allow missing exposure at baseline
  # this gives us greater confidence that we recover a incident effect
  filter(wave == baseline_wave, !is.na(!!sym(name_exposure))) |>
  pull(id)

# n eligible
n_participants <- length(ids_baseline)

# make pretty number
n_participants = margot::pretty_number(n_participants)

# save
here_save(n_participants, "n_participants")

cli::cli_h1("set eligibility criteria for baseline cohort ✓")
```

```
# +-------------------------+
# |END OPTIONALLY MODIFY SEC.|
# +-------------------------+
# +-------------------------+
# |       END ALERT         |
# +-------------------------+

# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# |    MODIFY THIS SECTION   |
# +-------------------------+
# plot distribution to help with cutpoint decision
dat_long_exposure <- dat_long_1 |> filter(wave %in% exposure_waves)

min_value <- min(dat_long_exposure$hours_social_media, na.rm = TRUE)
min_value

max_value <- max(dat_long_exposure$hours_social_media, na.rm = TRUE)
max_value
median_value <- median(dat_long_exposure$hours_social_media, na.rm = TRUE)

# view median
median_value
# define cutpoints for graph ----------------------------------------------

# define cutpoints *-- these can be adjusted --*

# to use later in positivity graph in manuscript
lower_cut <- 0
upper_cut <- 14
threshold <- '>' # if upper
inverse_threshold <- '<='
scale_range = 'scale range 0-168'

cut_points = c(lower_cut, upper_cut)


# save for manuscript
here_save(lower_cut, "lower_cut")
here_save(upper_cut, "upper_cut")
here_save(threshold, "threshold")
here_save(inverse_threshold, "inverse_threshold")
here_save(scale_range, "scale_range")


cli::cli_h1("set thresholds for binary variable (if variable is continuous) ✓")
```

```
hist(dat_long_exposure$hours_social_media)
cut_points

cutpoint_inclusive_value  <-  "lower"
# make graph
graph_cut <- margot::margot_plot_categorical(
  dat_long_exposure,
  col_name         = name_exposure,
  sd_multipliers = c(-1, 1), # select to suit
  # either use n_divisions for equal-sized groups:
  # n_divisions      = 2,
  # or use custom_breaks for specific values:
   custom_breaks  = c(lower_cut, upper_cut),  # ** adjust as needed **
  # could be "lower", no difference in this case, as no one == 4
  cutpoint_inclusive = cutpoint_inclusive_value,
  show_mean        = TRUE,
  # show_median      = TRUE,
  show_sd          = TRUE,
  binwidth = 1
)
print(graph_cut)

# save your graph
margot::here_save(graph_cut, "graph_cut", push_mods)

# create binary exposure variable based on chosen cutpoint
dat_long_2 <- margot::create_ordered_variable(
  dat_long_1,
  var_name           = name_exposure,
  custom_breaks  = c(lower_cut, upper_cut),  # ** adjust as needed **
  cutpoint_inclusive = cutpoint_inclusive_value,
)


cli::cli_h1("created binary variable (if variable is continuous) ✓")


# +-------------------------+
# |   END MODIFY SECTION    |
# +-------------------------+
# +-------------------------+
# |       END ALERT         |
# +-------------------------+



# +-------------------------+
# |      DO NOT ALTER       |
# +-------------------------+
```

```
# process binary variables and log-transform ---------------------------------
# convert binary factors to 0/1 format
dat_long_3 <- margot::margot_process_binary_vars(dat_long_2)

# log-transform hours and income variables: tables for analysis (only logged versions of vars)
dat_long_final <- margot::margot_log_transform_vars(
  dat_long_3,
  vars          = c(starts_with("hours_"), "household_inc"), # **--- think about this ---***
  prefix        = "log_",
  keep_original   = FALSE,
  exceptions = name_exposure # omit original variables#  **--- think about this ---***
) |>
  # select only variables needed for analysis
  select(all_of(c(baseline_vars, exposure_var, outcome_vars, "id", "wave", "year_measured",
"sample_weights"))) |>
  droplevels()

name_exposure


# check missing data ----------------------------------------------------
# this is crucial to understand potential biases
missing_summary <- naniar::miss_var_summary(dat_long_final)
print(missing_summary)
margot::here_save(missing_summary, "missing_summary", push_mods)



# visualise missing data pattern
# ** -- takes a while to render **
vis_miss <- naniar::vis_miss(dat_long_final, warn_large_data = FALSE)
print(vis_miss)
margot::here_save(vis_miss, "vis_miss", push_mods)

# calculate percentage of missing data at baseline
dat_baseline_pct <- dat_long_final |> filter(wave == baseline_wave)
percent_missing_baseline <- naniar::pct_miss(dat_baseline_pct)
margot::here_save(percent_missing_baseline, "percent_missing_baseline", push_mods)

# save prepared dataset for next stage ----------------------------------------
margot::here_save(dat_long_final, "dat_long_final", push_mods)


cli::cli_h1("made and saved final long data set for further processign in script 02 ✓")


# +-------------------------+
# |    END DO NOT ALTER     |
# +-------------------------+
```

```
# check positivity ----------------------------------------------------

# +-------------------------+
# |         ALERT           |
# +-------------------------+
# +-------------------------+
# |    MODIFY THIS SECTION   |
# +-------------------------+


# check
threshold # defined above
upper_cut # defined above
name_exposure # defined above



# create transition matrices to check positivity ----------------------------
# this helps assess whether there are sufficient observations in all exposure states
dt_positivity <- dat_long_final |>
  filter(wave %in% c(baseline_wave, exposure_waves)) |>
  select(!!sym(name_exposure), id, wave) |>
  mutate(exposure = round(as.numeric(!!sym(name_exposure)), 0)) |>
  # create binary exposure based on cutpoint
  mutate(exposure_binary = ifelse(exposure > upper_cut, 1, 0)) |> # check
  ## *-- modify this --*
  mutate(wave = as.numeric(wave) -1 )

# create transition tables
transition_tables <- margot::margot_transition_table(
  dt_positivity,
  state_var = "exposure",
  id_var = "id",
  waves = c(0, 1),
  wave_var = "wave",
  table_name = "transition_table"
)

# check
print(transition_tables$tables[[1]])

# save
margot::here_save(transition_tables, "transition_tables", push_mods)

# create binary transition tables
transition_tables_binary <- margot::margot_transition_table(
  dt_positivity,
  state_var = "exposure_binary",
  id_var = "id",
```

```
  waves = c(0, 1),
  wave_var = "wave",
  table_name = "transition_table_binary"
)

# check
print(transition_tables_binary$tables[[1]])

# save
margot::here_save(transition_tables_binary, "transition_tables_binary", push_mods)

# +-------------------------+
# |       END ALERT         |
# +-------------------------+

# create tables ------------------------------------------------------------
# baseline variable labels
var_labels_baseline <- list(
  # demographics
  "age" = "Age",
  "born_nz_binary" = "Born in New Zealand",
  "education_level_coarsen" = "Education Level",
  "employed_binary" = "Employed",
  "eth_cat" = "Ethnicity",
  "male_binary" = "Male",
  "not_heterosexual_binary" = "Non-heterosexual",
  "parent_binary" = "Parent",
  "partner_binary" = "Has Partner",
  "rural_gch_2018_l" = "Rural Classification",
  "sample_frame_opt_in_binary" = "Sample Frame Opt-In",

  # economic & social status
  "household_inc" = "Household Income",
  "log_household_inc" = "Log Household Income",
  "nz_dep2018" = "NZ Deprivation Index",
  "nzsei_13_l" = "Occupational Prestige Index",
  "household_inc" = "Household Income",


  # personality traits
  "agreeableness" = "Agreeableness",
  "conscientiousness" = "Conscientiousness",
  "neuroticism" = "Neuroticism",
  "openness" = "Openness",

  # beliefs & attitudes
  "political_conservative" = "Political Conservatism",
  "religion_identification_level" = "Religious Identification",

  # health behaviors
```

```
  "alcohol_frequency" = "Alcohol Frequency",
  "alcohol_intensity" = "Alcohol Intensity",
  "hlth_disability_binary" = "Disability Status",
  "smoker_binary" = "Smoker",
  "hours_exercise" = "Hours of Exercise",


  # time use
  "hours_children" = "Hours with Children",
  "hours_commute" = "Hours Commuting",
  "hours_exercise" = "Hours Exercising",
  "hours_housework" = "Hours on Housework",
  "log_hours_children" = "Log Hours with Children",
  "log_hours_commute" = "Log Hours Commuting",
  "log_hours_exercise" = "Log Hours Exercising",
  "log_hours_housework" = "Log Hours on Housework"
)
here_save(var_labels_baseline, "var_labels_baseline")

# outcome variable labels, organized by domain
# reivew your outcomes make sure they appear on the list below
# comment out what you do not need
outcome_vars

# get names
var_labels_outcomes <- list(
  "kessler_latent_anxiety" = "Anxiety",
  "kessler_latent_depression" = "Depression",
  "rumination" = "Rumination",
  "bodysat" = "Body Satisfaction",
  "self_esteem" = "Self Esteem",
  "lifesat" = "Life Satisfaction",
  "pwi" = "Personal Well-being Index",
  "perfectionism" = "Perfectionism",
  "lifemeaning" = "Life Meaning"
)

# save for manuscript
here_save(var_labels_outcomes, "var_labels_outcomes")




# save all variable translations
var_labels_measures <- c(var_labels_baseline, var_labels_exposure, var_labels_outcomes)
var_labels_measures

# save for manuscript
here_save(var_labels_measures, "var_labels_measures")
```

```
# +-------------------------+
# |   END MODIFY SECTION    |
# +-------------------------+

# +-------------------------+
# |      DO NOT ALTER       |
# +-------------------------+
# tables ------------------------------------------------------------------
# create baseline characteristics table
dat_baseline = dat_long_final |>
  filter(wave %in% c(baseline_wave)) |>
  mutate(
    male_binary = factor(male_binary),
    parent_binary = factor(parent_binary),
    smoker_binary = factor(smoker_binary),
    born_nz_binary = factor(born_nz_binary),
    employed_binary = factor(employed_binary),
    not_heterosexual_binary = factor(not_heterosexual_binary),
    sample_frame_opt_in_binary = factor(sample_frame_opt_in_binary)
  )


# +-------------------------+
# |         ALERT           |
# +-------------------------+

# save sample weights from baseline wave
# save sample weights
t0_sample_weights <- dat_baseline$sample_weights
here_save(t0_sample_weights, "t0_sample_weights")

# +-------------------------+
# |       END ALERT         |
# +-------------------------+


# make baseline table -----------------------------------------------------

baseline_table <- margot::margot_make_tables(
  data = dat_baseline,
  vars = baseline_vars,
  by = "wave",
  labels = var_labels_baseline,
  table1_opts = list(overall = FALSE, transpose = FALSE),
  format = "markdown"
)
print(baseline_table)
margot::here_save(baseline_table, "baseline_table", push_mods)

# create exposure table by wave
```

```
exposure_table <- margot::margot_make_tables(
  data = dat_long_final |> filter(wave %in% c(baseline_wave, exposure_waves)),
  vars = exposure_var,
  by = "wave",
  labels = var_labels_exposure,
  factor_vars = exposure_var_binary,
  table1_opts = list(overall = FALSE, transpose = FALSE),
  format = "markdown"
)
print(exposure_table)
margot::here_save(exposure_table, "exposure_table", push_mods)

# create outcomes table by wave
outcomes_table <- margot::margot_make_tables(
  data = dat_long_final |> filter(wave %in% c(baseline_wave, outcome_wave)),
  vars = outcome_vars,
  by = "wave",
  labels = var_labels_outcomes,
  format = "markdown"
)
print(outcomes_table)
margot::here_save(outcomes_table, "outcomes_table", push_mods)

# +--------------------------+
# |    END DO NOT ALTER     |
# +--------------------------+


# +--------------------------+
# |    END                  |
# +--------------------------+


# note: completed data preparation step -------------------------------------
# you're now ready for the next steps:
# 1. creating wide-format dataset for analysis
# 2. applying causal inference methods
# 3. conducting sensitivity analyses

# key decisions summary:
# exposure variable: extraversion
# study waves: baseline (2018), exposure (2019), outcome (2020)
# baseline covariates: demographics, traits, health measures (excluding exposure)
# outcomes: health, psychological, wellbeing, and social variables
# binary cutpoint for exposure: here, 4 on the extraversion scale
# label names for tables
```

Script 2:

```
# script 2: causal workflow for estimating average treatment effects using margot
# may 2025
# questions: joseph.bulbulia@vuw.ac.nz

# +-------------------------+
# |      DO NOT ALTER       |
# +-------------------------+

# restart fresh session for a clean workspace
rstudioapi::restartSession()

# set seed for reproducibility
set.seed(123)

# libraries ------------------------------------------------------------
# essential library ----------------------------------------------------
if (!require(margot, quietly = TRUE)) {
  devtools::install_github("go-bayes/margot")
}




if (packageVersion("margot") < "1.0.47") {
  stop("please install margot >= 1.0.47 for this workflow\n
       run: devtools::install_github(\"go-bayes/margot\")
")
}


library(margot)


# load packages --------------------------------------------------------
# pacman will install missing packages automatically
if (!requireNamespace("pacman", quietly = TRUE)) install.packages("pacman")
pacman::p_load(
  tidyverse,       # data wrangling + plotting
  qs,              # fast data i/o
  here,            # project-relative file paths
  data.table,      # fast data manipulation
  fastDummies,     # dummy variable creation
  naniar,          # missing data handling
  skimr,           # summary statistics
  grf,             # machine learning forests
  kableExtra,      # tables
  ggplot2,         # graphs
```

```
  doParallel,      # parallel processing
  grf,             # causal forests
  janitor,         # variables names
  stringr,         # variable names
  patchwork,       # graphs
  table1,          # tables
  cli
)



# save paths ------------------------------------------------------------------
push_mods <- here::here("save_directory")

# read data
dat_long_final <- margot::here_read("dat_long_final")

# read baseline sample weights
t0_sample_weights <- margot::here_read("t0_sample_weights")

# read exposure
name_exposure <- margot::here_read("name_exposure")
name_exposure_binary = paste0(name_exposure, "_binary")
name_exposure_continuous = name_exposure

# read variables
baseline_vars <- margot::here_read("baseline_vars")
exposure_var <- margot::here_read("exposure_var")
outcome_vars <- margot::here_read("outcome_vars")
baseline_wave <- margot::here_read("baseline_wave")
exposure_waves <- margot::here_read("exposure_waves")
outcome_wave <- margot::here_read("outcome_wave")

# define continuous columns to keep
continuous_columns_keep <- c("t0_sample_weights")

# check is this the exposure variable that you want?
name_exposure_binary
name_exposure_continuous

# ordinal use
ordinal_columns <- c(
  "t0_education_level_coarsen",
  "t0_eth_cat",
  "t0_rural_gch_2018_l",
  "t0_gen_cohort",
  "t0_religion_bigger_denominations" # <- added for demonstration (optional)
)
```

```
# define wide variable names
t0_name_exposure_binary <- paste0("t0_", name_exposure_binary)
t0_name_exposure_binary

# make exposure names (continuous not genreally used)
t1_name_exposure_binary <- paste0("t1_", name_exposure_binary)
t1_name_exposure_binary

# treatments (continuous verion)
t0_name_exposure <- paste0("t0_", name_exposure_continuous)
t1_name_exposure <- paste0("t1_", name_exposure_continuous)
t0_name_exposure_continuous <- paste0("t0_", name_exposure)
t1_name_exposure_continuous <- paste0("t1_", name_exposure)

# raw outcomes
# read health outcomes
outcome_vars <- here_read("outcome_vars")
t2_outcome_z <- paste0("t2_", outcome_vars, "_z")

# view
t2_outcome_z

# check
str(dat_long_final)

# check
naniar::gg_miss_var(dat_long_final)

# impute data ----------------------------------------------------------

# define cols we will not standardise
continuous_columns_keep <- c("t0_sample_weights")

# remove sample weights
dat_long_final_2 <- dat_long_final |> select(-sample_weights)

# prepare data for analysis ----------------------
dat_long_final_2 <- margot::remove_numeric_attributes(dat_long_final_2)
# wide data
df_wide <- margot_wide_machine(
  dat_long_final,
  id = "id",
  wave = "wave",
  baseline_vars,
  exposure_var = exposure_var,
  outcome_vars,
  confounder_vars = NULL,
  imputation_method = "none",
  include_exposure_var_baseline = TRUE,
  include_outcome_vars_baseline = TRUE,
```

```
  extend_baseline = FALSE,
  include_na_indicators = FALSE
)

# check
colnames(df_wide)

# return sample weights
df_wide$t0_sample_weights <-  t0_sample_weights

# save
margot::here_save(df_wide, "df_wide")

#df_wide <- margot::here_read("df_wide")
naniar::vis_miss(df_wide, warn_large_data = FALSE)

# view
glimpse(df_wide)


# order data with missingness assigned to work with grf and lmtp
# if any outcome is censored all are censored
# create version for model reports

# check
colnames(df_wide)


# made data wide in correct format
# ** ignore warning ***
df_wide_encoded  <- margot::margot_process_longitudinal_data_wider(
  df_wide,
  ordinal_columns = ordinal_columns, #<- make sure all ordinal columns have been identified
  continuous_columns_keep = continuous_columns_keep,
  not_lost_in_following_wave = "not_lost_following_wave",
  lost_in_following_wave = "lost_following_wave",
  remove_selected_columns = TRUE,
  exposure_var = exposure_var,
  scale_continuous = TRUE
)


# check
colnames(df_wide_encoded)

# check
table(df_wide_encoded$t0_not_lost_following_wave)

# make the binary variable numeric
df_wide_encoded[[t0_name_exposure_binary]] <-
```

```
  as.numeric(df_wide_encoded[[t0_name_exposure_binary]]) - 1
df_wide_encoded[[t1_name_exposure_binary]] <-
  as.numeric(df_wide_encoded[[t1_name_exposure_binary]]) - 1

# view
df_wide_encoded[[t0_name_exposure_binary]]
df_wide_encoded[[t1_name_exposure_binary]]

# 1. ensure both binaries only take values 0 or 1 (ignore NA)
stopifnot(all(df_wide_encoded[[t0_name_exposure_binary]][!is.na(df_wide_encoded[[t0_na
me_exposure_binary]])] %in% 0:1),

all(df_wide_encoded[[t1_name_exposure_binary]][!is.na(df_wide_encoded[[t1_name_expos
ure_binary]])] %in% 0:1))

# 2. ensure NA-patterns match between t1_exposure and t0_lost flag
# count n-as in t1 exposure
n_na_t1 <- sum(is.na(df_wide_encoded[[t1_name_exposure_binary]]))

# count how many were lost at t0
n_lost_t0 <- sum(df_wide_encoded$t0_lost_following_wave == 1, na.rm = TRUE)

# print them for inspection
message("NAs in ", t1_name_exposure_binary, ": ", n_na_t1)
message("t0_lost_following_wave == 1: ", n_lost_t0)

# stop if they don't match
stopifnot(n_na_t1 == n_lost_t0)

# 3. ensure if t1 is non-NA then subject was not lost at t0
stopifnot(all(is.na(df_wide_encoded[[t1_name_exposure_binary]]) |
              df_wide_encoded[["t0_not_lost_following_wave"]] == 1))

# view
glimpse(df_wide_encoded)

table(is.na( df_wide_encoded$t1_hours_social_media_binary))

# check will take a while
#naniar::vis_miss(df_wide_encoded, warn_large_data = FALSE)
naniar::gg_miss_var(df_wide_encoded)


#save data
here_save(df_wide_encoded, "df_wide_encoded")

# new weights approach ----------------------------------------------------


# panel attrition workflow using grf (two-stage IPCW + design weights)
```

```
# ------------------------------------------------------------------------
# builds weights in two stages:
#   w0 : baseline -> t1  (baseline covariates)
#   w1 : t1 survivors -> t2  (baseline + time-1 exposure)
# final weight = t0_sample_weights × w0 × w1, then trimmed & normalised.
# ------------------------------------------------------------------------

# ── 0 setup ──────────────────────────────────────────────────────────────

library(tidyverse)       # wrangling
library(glue)            # strings
library(grf)             # forests
library(cli)             # progress


set.seed(123)

# ------------------------------------------------------------------------
# 1 import full, unfiltered baseline file
# ------------------------------------------------------------------------

df <- margot::here_read("df_wide_encoded")
cli::cli_alert_info(glue("{nrow(df)} rows × {ncol(df)} columns loaded"))


# ------------------------------------------------------------------------
# 2 stage-0 censoring: dropout between t0 → t1
# ------------------------------------------------------------------------

baseline_covars <- df %>%
  select(starts_with("t0_"), -ends_with("_lost"), -ends_with("lost_following_wave"), -
ends_with("_weights")) %>%
  colnames() %>% sort()

# select your baseline vars and coerce to numeric
num_dat <- df %>%
  select(all_of(baseline_covars)) %>%
  mutate(across(everything(), as.numeric))

# build a true numeric matrix
X0 <- as.matrix(num_dat)

# make factor
D0 <- factor(df$t0_lost_following_wave, levels = c(0, 1))   # 0 = stayed, 1 = lost

cli::cli_h1("stage 0: probability forest for baseline dropout …")

# then fit
```

```
pf0 <- probability_forest(X0, D0)
P0  <- predict(pf0, X0)$pred[, 2]              # P(dropout by t1)
w0  <- ifelse(D0 == 1, 0, 1 / (1 - P0))        # IPCW for stage 0
df$w0 <- w0

# ------------------------------------------------------------------------
# 3 stage-1 censoring: dropout between t1 → t2 (baseline + exposure)
# ------------------------------------------------------------------------

exposure_var <- t1_name_exposure_binary       # ← binary exposure variable name

# filter out those lost (already weighted for censoring)
df1 <- df %>% filter(t0_lost_following_wave == 0)

# filter to those at risk in stage-1
cen1_data <- df %>%
  filter(t0_lost_following_wave == 0,
         !is.na(.data[[exposure_var]]))

# coerce baseline covars + exposure all at once
X1_num <- cen1_data %>%
  # convert every t0_… and the exposure to numeric
  mutate(across(all_of(c(baseline_covars, exposure_var)), as.numeric)) %>%
  # now select in the order you want
  select(all_of(baseline_covars), all_of(exposure_var))

# build numeric matrix
X1 <- as.matrix(X1_num)
colnames(X1)[ncol(X1)] <- exposure_var

D1 <- factor(cen1_data$t1_lost_following_wave, levels = c(0, 1))

cli::cli_h1("stage 1: probability forest for second-wave dropout …")
pf1 <- probability_forest(X1, D1)
P1  <- predict(pf1, X1)$pred[, 2]
w1  <- ifelse(D1 == 1, 0, 1 / (1 - P1))

# map w1 back to df1 (rows with NA exposure get weight 0)
df1$w1 <- 0
df1$w1[match(cen1_data$id, df1$id)] <- w1

# ------------------------------------------------------------------------
# 4 combine design × IPCW weights
# ------------------------------------------------------------------------

# bring forward w0 for the matching rows (safe join)
w0_vec <- df$w0[match(df1$id, df$id)]

# combined weight before trim / normalise
raw_w <- df1$t0_sample_weights * w0_vec * df1$w1
```

```
df1$raw_weight <- raw_w

# trim + normalise (exclude NA & zeros)
pos <- raw_w[!is.na(raw_w) & raw_w > 0]

lb  <- quantile(pos, 0.00, na.rm = TRUE)
ub  <- quantile(pos, 0.99, na.rm = TRUE)

trimmed <- pmin(pmax(raw_w, lb), ub)
normalised <- trimmed / mean(trimmed, na.rm = TRUE)

df1$combo_weights <- normalised <- trimmed / mean(trimmed)

df1$combo_weights <- normalised

hist(df1$combo_weights[df1$t1_lost_following_wave == 0],
     main = "combined weights (observed)", xlab = "weight")

# ------------------------------------------------------------------------
# 5 analysis set: observed through t2 (not censored at either stage)
# ------------------------------------------------------------------------

df_analysis <- df1 %>%
  filter(t1_lost_following_wave == 0) %>%
  droplevels()

margot::here_save(df_analysis, "df_analysis_weighted_two_stage")

cli::cli_alert_success(glue("analysis sample: {nrow(df_analysis)} obs"))

# TEST DO NOT UNCOMMENT
# ------------------------------------------------------------------------
# 6 causal forest (edit outcome var if needed)
# ------------------------------------------------------------------------
#
# outcome_var <- "t2_kessler_latent_depression_z"   # ← edit
#
# Y <- df_analysis[[outcome_var]]
# W <- df_analysis[[exposure_var]]
# X <- as.matrix(df_analysis[, baseline_covars])
#
# cf <- causal_forest(
#   X, Y, W,
#   sample.weights = df_analysis$combo_weights,
#   num.trees      = 2000
# )
#
# print(average_treatment_effect(cf))
# margot::here_save(cf,          "cf_ipcw_two_stage")
```

```
# ------------------------------------------------------------------------
# 7 save objects
# ------------------------------------------------------------------------


cli::cli_h1("two-stage IPCW workflow complete ✓")


# # maintain workflow
E <- baseline_covars

here_save(E, "E")
length(E)
colnames(df_analysis)


cli::cli_h1("naming convention matcheds `grf` ✓")


# arrange
df_grf <- df_analysis |>
  relocate(ends_with("_weights"), .before = starts_with("t0_")) |>
  relocate(ends_with("_weight"), .before = ends_with("_weights")) |>
  relocate(starts_with("t0_"), .before = starts_with("t1_")) |>
  relocate(starts_with("t1_"), .before = starts_with("t2_")) |>
  relocate("t0_not_lost_following_wave", .before = starts_with("t1_")) |>
  relocate(all_of(t1_name_exposure_binary), .before = starts_with("t2_")) |>
  droplevels()

colnames(df_grf)


# +--------------------------+
# |          ALERT           |
# +--------------------------+
# make sure to do this
# save final data
margot::here_save(df_grf, "df_grf")

cli::cli_h1("saved data `df_grf` for models ✓")


# +--------------------------+
# |       END ALERT          |
# +--------------------------+


# check final dataset
colnames(df_grf)
```

```
# visualise missing
# should have no missing in t1 and t2 variables
# handled by IPCW
# make final missing data graph
missing_final_data_plot <- naniar::vis_miss(df_grf, warn_large_data = FALSE)
missing_final_data_plot

# save plot
margot_save_png(missing_final_data_plot, prefix = "missing_final_data")

# checks
colnames(df_grf)
str(df_grf)

# check exposures
table(df_grf[[t1_name_exposure_binary]])

# check
hist(df_grf$combo_weights)

# calculate summary statistics
t0_weight_summary <- summary(df_wide_encoded)

# check
glimpse(df_grf$combo_weights)

# visualise weight distributions
hist(df_grf$combo_weights, main = "t0_stabalised weights", xlab = "Weight")

# check n
n_observed_grf <- nrow(df_grf)

# view
n_observed_grf

# save
margot::here_save(n_observed_grf, "n_observed_grf")



# +-------------------------+
# |    END DO NOT ALTER     |
# +-------------------------+


# +-------------------------+
# |    END                  |
# +-------------------------+

# this is just for your interest ----------------------------------------
```

```
# not used in final manuscript
# FOR INTEREESTS
# inspect propensity scores ---------------------------------------------
# get data
# df_grf <- here_read('df_grf')
#
# # assign weights var name
# weights_var_name = "t0_adjusted_weights"#
# # baseline covariates  # E already exists and is defined
# E
#
# # must be a data frame, no NA in exposure
#
# # df_grf is a data frame - we must process this data frame in several steps
# # user to specify which columns are outcomes, default to 'starts_with("t2_")'
# df_propensity_org <- df_grf |> select(!starts_with("t2_"))
#
# # Remove NAs and print message that this has been done
# df_propensity <- df_propensity_org |> drop_na() |> droplevels()
#
# # E_propensity_names
# # first run model for baseline propensity if this is selected.  The default should be to not
select it.
# propensity_model_and_plots <- margot_propensity_model_and_plots(
#   df_propensity = df_propensity,
#   exposure_variable = t1_name_exposure_binary,
#   baseline_vars = E,
#   weights_var_name = weights_var_name,
#   estimand = "ATE",
#   method = "ebal",
#   focal = NULL
# )
#
# # visualise
# summary(propensity_model_and_plots$match_propensity)
#
# # key plot
# propensity_model_and_plots$love_plot
#
# # other plots
# propensity_model_and_plots$summary_plot
# propensity_model_and_plots$balance_table
# propensity_model_and_plots$diagnostics
#
#
# # check size
# size_bytes <- object.size(propensity_model_and_plots)
# print(size_bytes, units = "auto") # Mb
#
# # use qs to save only if you have space
```

```
# here_save_qs(propensity_model_and_plots,
#              "propensity_model_and_plots",
#              push_mods)
```

Script 3:

```
# script 3: causal workflow for estimating average treatment effects using margot
# may 2025
# questions: joseph.bulbulia@vuw.ac.nz

# +-------------------------+
# |      DO NOT ALTER       |
# +-------------------------+

# restart fresh session

rstudioapi::restartSession()


# reproducibility ---------------------------------------------------------

# choose number
my_seed = 123
set.seed(my_seed)


# essential library -------------------------------------------------------
if (!require(margot, quietly = TRUE)) {
  devtools::install_github("go-bayes/margot")
  library(margot)
}

# min version of margot
if (packageVersion("margot") < "1.0.52") {
  stop(
    "please install margot >= 1.0.52 for this workflow\n
       run: devtools::install_github('go-bayes/margot')
"
  )
}

# call library
library("margot")

# check package version
packageVersion(pkg = "margot")


# load libraries ----------------------------------------------------------
# pacman will install missing packages automatically
```

```
pacman::p_load(
  tidyverse,
  # data wrangling + plotting
  qs,
  here,# project-relative file paths
  data.table,# fast data manipulation
  fastDummies,# dummy variable creation
  naniar,# missing data handling
  skimr,# summary statistics
  grf,
  ranger,
  doParallel,
  kableExtra,
  ggplot2 ,
  rlang ,
  purrr ,
  patchwork,
  janitor,  # nice labels
  glue,
  cli,
  future,
  crayon,
  glue,
  stringr,
  future,
  furrr
)

# directory path configuration ------------------------------------------------
# save path (customise for your own computer) ----------------------------
push_mods <- here::here("save_directory")

# read original data (for plots) ----------------------------------------
original_df <- margot::here_read("df_wide", push_mods)


# plot title ------------------------------------------------------------
title_binary = "Effects of Social Media on Well-being"
filename_prefix = "grf_socialmedia_wb"

# for manuscript later
margot::here_save(title_binary, "title_binary")

# import names ----------------------------------------------------------
name_exposure <- margot::here_read("name_exposure")
name_exposure

# make exposure names
t1_name_exposure_binary <- paste0("t1_", name_exposure, "_binary")
```

```
# check exposure name
t1_name_exposure_binary

# read outcome vars
outcome_vars <- margot::here_read("outcome_vars")

# read and sort outcome variables ----------------------------------------
# we do this by domain: health, psych, present, life, social
read_and_sort <- function(key) {
  raw  <- margot::here_read(key, push_mods)
  vars <- paste0("t2_", raw, "_z")
  sort(vars)
}
t2_outcome_z  <- read_and_sort("outcome_vars")

# view
t2_outcome_z


# +-------------------------+
# |    END DO NOT ALTER     |
# +-------------------------+


# +-------------------------+
# |   MODIFY THIS SECTION   |
# +-------------------------+


# define names for titles ------------------------------------------------

nice_exposure_name =  stringr::str_to_sentence(name_exposure)
nice_outcome_name = "Wellbeing"
title = glue::glue("Effect of Social Media on Well-Being")
title
# save for final report
here_save(title, "title")



# combine outcomes -------------------------------------------------------
# check outcome vars and make labels for graphs/tables
outcome_vars

label_mapping_all <- list(
  "t2_kessler_latent_anxiety_z" = "Anxiety",
  "t2_kessler_latent_depression_z" = "Depression",
  "t2_rumination_z" = "Rumination",
  "t2_bodysat_z" = "Body Satisfaction",
  "t2_self_esteem_z" = "Self Esteem",
```

```
  "t2_lifesat_z" = "Life Satisfaction",
  "t2_pwi_z" = "Personal Well-being Index",
  "t2_perfectionism_z" = "Perfectionism",
  "t2_lifemeaning_z" = "Life Meaning"
)

# save
here_save(label_mapping_all, "label_mapping_all")

# check
label_mapping_all

cli::cli_h1("created and saved label_mapping for use in graphs/tables ✓")


# make options ------------------------------------------------------------
# titles
ate_title = "ATE Effects of Social Media on Well-Being"
subtitle = ""
filename_prefix = "final_report"
#
here_save(ate_title, "ate_title")
here_save(filename_prefix, "filename_prefix")

# settings
x_offset = -.25
x_lim_lo = -.25
x_lim_hi = .25

# defaults for ate plots
base_defaults_binary <- list(
  type = "RD",
  title = ate_title,
  e_val_bound_threshold = 1.2,
  colors = c(
    "positive" = "#E69F00",
    "not reliable" = "grey50",
    "negative" = "#56B4E9"
  ),
  x_offset = x_offset,
  # will be set based on type
  x_lim_lo = x_lim_lo,
  # will be set based on type
  x_lim_hi = x_lim_hi,
  text_size = 8,
  linewidth = 0.75,
  estimate_scale = 1,
  base_size = 18,
  point_size = 4,
  title_size = 19,
```

```
  subtitle_size = 16,
  legend_text_size = 10,
  legend_title_size = 10,
  include_coefficients = FALSE
)

# save

# health graph options
outcomes_options_all <- margot_plot_create_options(
  title = subtitle,
  base_defaults = base_defaults_binary,
  subtitle = subtitle,
  filename_prefix = filename_prefix
)


# policy tree graph settings ----------------------------------------------
decision_tree_defaults <- list(
  span_ratio      = .3,
  text_size       = 3.8,
  y_padding       = 0.25,
  edge_label_offset = .002,
  border_size     = .05
)
policy_tree_defaults <- list(
  point_alpha      = .5,
  title_size       = 12,
  subtitle_size    = 12,
  axis_title_size  = 12,
  legend_title_size = 12,
  split_line_color  = "red",
  split_line_alpha  = .8,
  split_label_color = "red",
  list(split_label_nudge_factor = 0.007)
)

# +-------------------------+
# |   END MODIFY SECTION    |
# +-------------------------+

# +-----------------------------------------+
# |      DO NOT ALTER  (except where noted)     |
# +-----------------------------------------+
# load GRF data and prepare inputs ----------------------------------------
df_grf <- margot::here_read('df_grf', push_mods)
E      <- margot::here_read('E', push_mods)

# check exposure binary
```

```
stopifnot(all(df_grf[[t1_name_exposure_binary]][!is.na(df_grf[[t1_name_exposure_binary]])
] %in% 0:1))
# set exposure and weights

W       <- as.vector(df_grf[[t1_name_exposure_binary]]) # note it is the processed weights for
attrition "t1"

# old workflow
# weights <- df_grf$t1_adjusted_weights

# new weights workflow, use "combo_weights" -- see revised script 2
weights <- df_grf$combo_weights

hist(weights) # quick check for extreme weights
# select covariates and drop numeric attributes
X <- margot::remove_numeric_attributes(df_grf[E])

# set model defaults ----------------------------------------------------
grf_defaults <- list(seed = 123,
                     stabilize.splits = TRUE,
                     num.trees = 2000)


# causal forest model ---------------------------------------------------------

# +--------------------------+
# |         ALERT            |
# +--------------------------+

# !!!! THIS WILL TAKE TIME  !!!!!
# **----- COMMENT OUT AFTER YOU RUN TO AVOID RUNNING MORE THAN
ONCE -----**


models_binary <- margot_causal_forest(
  # <- could be 'margot_causal_forest_parrallel()' if you have a powerful computer
  data = df_grf,
  outcome_vars = t2_outcome_z,
  covariates = X,
  W = W,
  weights = weights,
  grf_defaults = grf_defaults,
  top_n_vars = 15,
  #<- can be modified but will affect run times
  save_models = TRUE,
  save_data = TRUE,
  train_proportion = 0.7
)
```

```
# +-------------------------+
# |         ALERT           |
# +-------------------------+
# !!!! THIS WILL TAKE TIME  !!!!!
# save model
margot::here_save_qs(models_binary, "models_binary", push_mods)
# +-------------------------+
# |       END ALERT         |
# +-------------------------+

cli::cli_h1("causal forest model completed and saved ✓")

# read results ------------------------------------------------------------
# if you save models you do not need to re-run them


# +-------------------------+
# |         ALERT           |
# +-------------------------+
# !!!! THIS WILL TAKE TIME  !!!!!
models_binary <- margot::here_read_qs("models_binary", push_mods)
# +-------------------------+
# |       END ALERT         |
# +-------------------------+

# count models by category
# just a check
cat("Number of original models:\n",
    length(models_binary$results),
    "\n")

# look at the hat(tau(x))
plot_tau_hats <- margot_plot_tau(models_binary, label_mapping = label_mapping_all)

# view
plot_tau_hats

# make ate plots ----------------------------------------------------------
#   ********** NEW - CORRECTION FOR FAMILY-WISE ERROR **********
# then pass to the results
ate_results <- margot_plot(
  models_binary$combined_table,
  # <- now pass the corrected results.
  options = outcomes_options_all,
  label_mapping = label_mapping_all,
  include_coefficients = FALSE,
  save_output = FALSE,
  order = "evaluebound_asc",
  original_df = original_df,
  e_val_bound_threshold = 1.2,
```

```
  rename_ate = TRUE,
  adjust = "bonferroni",  #<- new
  alpha = 0.05 # <- new
)


# view
cat(ate_results$interpretation)

# check
ate_results$plot

# interpretation
cat(ate_results$interpretation)

# save
here_save_qs(ate_results, "ate_results", push_mods)


# make markdown tables (to be imported into the manuscript)
margot_bind_tables_markdown <- margot_bind_tables(
  ate_results$transformed_table,
  #list(all_models$combined_table),
  sort_E_val_bound = "desc",
  e_val_bound_threshold = 1.2,
  # ← choose threshold
  highlight_color = NULL,
  bold = TRUE,
  rename_cols = TRUE,
  col_renames = list("E-Value" = "E_Value", "E-Value bound" = "E_Val_bound"),
  rename_ate = TRUE,
  threshold_col = "E_Val_bound",
  output_format = "markdown",
  kbl_args = list(
    booktabs = TRUE,
    caption = NULL,
    align = NULL
  )
)

# view markdown table
margot_bind_tables_markdown

# save for publication
here_save(margot_bind_tables_markdown, "margot_bind_tables_markdown")


# evaluate models ------------------------------------------------------
# trim models if extreme propensity scores dominate
# diag_tbl_98 <- margot_inspect_qini(models_binary,
```

```
#                                 propensity_bounds = c(0.01, 0.99))



# +-------------------------+
# |    END DO NOT ALTER     |
# +-------------------------+



# +-------------------------+
# |   MODIFY THIS SECTION   |
# +-------------------------+


# FLIPPING OUTCOMES  ----------------------------------------------------

# note that the meaning of a heterogeneity will vary depending on our interests.
# typically we are interested in whether an exposure improves life, and whether there is variability (aka HTE) in degrees of improvement.
# in this case we must take negative outcomes and "flip" them -- recalculating the policy trees and qini curves for each
# for example if the outcome is depression, then by flipping depression we better understand how the exposure *reduces* depression.
# what if the exposure is harmful? say what if we are interested in the effect of depression on wellbeing? In that case, we might
# want to "flip" the positive outcomes. That is, we might want to understand for whom a negative exposure is extra harmful.
# here we imagine that extroversion is generally positive in its effects, and so we "flip" the negative outcomes.
# if you were interested in a negative exposure, say "neuroticism" then you would probably want to flip the positive outcomes.
# note there are further questions we might ask. We might consider who responds more 'weakly" to a negative exposure (or perhaps to a positive exposure).
# Such a question could make sense if we had an exposure that was generally very strong.
# however, let's stay focussed on evaluating evaluating strong responders. We will flip the negative outcomes if we expect the exposure is positive,
# and flip the positive outcomes if we expect the exposure to be generally negative.
# if there is no natural "positive" or negative, then just make sure the valence of the outcomes aligns, so that all are oriented in the same
# direction if they have a valence.  if unsure, just ask for help!

# flipping models: outcomes we want to minimise given the exposure --------
# standard negative outcomes/  not used in this example
# flipping models: outcomes we want to minimise given the exposure --------
# standard negative outcomes/  not used in this example


# +-------------------------+
```

```
# |    MODIFY THIS          |
# +-------------------------+

# WHICH OUTCOMES -- if any ARE UNDESIREABLE?
flip_outcomes = c(
   # "t2_kessler_latent_anxiety_z",
   # "t2_kessler_latent_depression_z",
   # "t2_rumination_z",
   "t2_bodysat_z",
   "t2_self_esteem_z",
   "t2_lifesat_z",
   "t2_pwi_z",
   # "t2_perfectionism_z",
   "t2_lifemeaning_z"
 )

# save
here_save(flip_outcomes, "flip_outcomes")


# check
flip_outcomes
label_mapping_all


# +-------------------------+
# |   END MODIFY            |
# +-------------------------+

# get labels
flipped_names <- margot_get_labels(flip_outcomes, label_mapping_all)

# check
flipped_names

# save for publication
here_save(flipped_names, "flipped_names")

cli::cli_h1("flipped outcomes identified and names saved ✓")


# flip negatively oriented outcomes ------------------------------------

# +-------------------------+
# |      DO NOT ALTER       |
# +-------------------------+


# flip models using margot's function
```

```
#  *** this will take some time ***

# ** give it time **
# ** once run/ comment out **

# +-------------------------+
# |         ALERT           |
# +-------------------------+
# !!!! THIS WILL TAKE TIME  !!!!!
# can be margot_flip_forests_parrallel() if you have sufficient compute, set GB = something
less than your system RAM
models_binary_flipped_all <- margot_flip_forests(models_binary,
                                      flip_outcomes = flip_outcomes,
                                      recalc_policy = TRUE)

cli::cli_h1("flipped forest models completed ✓")


# !!!! THIS WILL TAKE TIME  !!!!!
# save
here_save_qs(models_binary_flipped_all,
           "models_binary_flipped_all",
           push_mods)


# +-------------------------+
# |         ALERT           |
# +-------------------------+
# !!!! THIS WILL TAKE TIME  !!!!!
# read back if needed
models_binary_flipped_all <- here_read_qs("models_binary_flipped_all", push_mods)


# this is a new function requires margot 1.0.48 or higher
label_mapping_all_flipped <- margot_reversed_labels(label_mapping_all, flip_outcomes)

# view
label_mapping_all_flipped


# save flipped labels
here_save(label_mapping_all_flipped, "label_mapping_all_flipped")

# +-------------------------+
# |       END ALERT         |
# +-------------------------+


# +-------------------------+
# |      DO NOT ALTER       |
```

```
# +-------------------------+


#
# ============================================================================
# SCRIPT:  HETEROGENEITY WORKFLOW
# PURPOSE: screen outcomes for heterogeneity, plot RATE & Qini curves,
#          fit shallow policy trees, and produce plain-language summaries.
# REQUIREMENTS:
#   • margot ≥ 1.0.52
#   • models_binary_flipped_all       – list returned by margot_causal_forest()
#   • original_df                     – raw data frame used in the forest
#   • label_mapping_all_flipped       – named vector of pretty labels
#   • flipped_names                   – vector of outcomes that were flipped
#   • decision_tree_defaults          – list of control parameters
#   • policy_tree_defaults            – list of control parameters
#   • push_mods                       – sub-folder for caches/outputs
#   • use 'models_binary', `label_mapping_all`, and set `flipped_names = ""` if no outcome flipped
#
# ============================================================================


# check package version early
stopifnot(utils::packageVersion("margot") >= "1.0.52")

# helper: quick kable printer ----------------------------------------------
print_rate <- function(tbl) {
  tbl |>
    mutate(across(where(is.numeric), \(x) round(x, 2))) |>
    kbl(format = "markdown")
}

# 1  SCREEN FOR HETEROGENEITY (RATE AUTOC + RATE Qini)  ----------------------

rate_results <- margot_rate(
  models       = models_binary_flipped_all,
  policy       = "treat_best",
  alpha        = 0.20,       # keep raw p < .20
  adjust       = "fdr",      # false-discovery-rate correction
  label_mapping = label_mapping_all_flipped
)

print_rate(rate_results$rate_autoc)
print_rate(rate_results$rate_qini)
# convert RATE numbers into plain-language text
rate_interp <- margot_interpret_rate(
  rate_results,
  flipped_outcomes      = flipped_names,
```

```
  adjust_positives_only = TRUE
)

cat(rate_interp$comparison, "\n")




cli_h2("Analysis ready for Appendix ✓")

# organise model names by evidence strength
model_groups <- list(
  autoc       = rate_interp$autoc_model_names,
  qini        = rate_interp$qini_model_names,
  either      = rate_interp$either_model_names,
  exploratory = rate_interp$not_excluded_either
)

# 2  PLOT RATE AUTOC CURVES ------------------------------------------------

autoc_plots <- margot_plot_rate_batch(
  models       = models_binary_flipped_all,
  save_plots    = FALSE,  # set TRUE to store .png files
  label_mapping = label_mapping_all_flipped,
  model_names   = model_groups$autoc
)

# inspect the first curve - note there may be more/none.
# if none, comment out
autoc_plots[[1]]

autoc_name_1 <- rate_results$rate_autoc$outcome[[1]]

# 3  QINI CURVES + GAIN INTERPRETATION ---------------------------------------
qini_results <- margot_policy(
  models_binary_flipped_all,
  save_plots         = FALSE,
  output_dir         = here::here(push_mods),
  decision_tree_args = decision_tree_args,
  policy_tree_args   = policy_tree_args,
  model_names        = names(models_binary_flipped_all$results),
  original_df        = original_df,
  label_mapping      = label_mapping_all_flipped,
  max_depth          = 2L,
  output_objects     = c("qini_plot", "diff_gain_summaries")
)
```

```
qini_gain <- margot_interpret_qini(
  qini_results,
  label_mapping = label_mapping_all_flipped
)

print_rate(qini_gain$summary_table)
cat(qini_gain$qini_explanation, "\n")

reliable_ids <- qini_gain$reliable_model_ids

# number of valid models for HTE
number_of_ids <- length(reliable_ids)
print( number_of_ids)

# (re-)compute plots only for models that passed Qini reliability
qini_results_valid <- margot_policy(
  models_binary_flipped_all,
  save_plots         = FALSE,
  output_dir         = here::here(push_mods),
  decision_tree_args = decision_tree_args,
  policy_tree_args   = policy_tree_args,
  model_names        = reliable_ids,
  original_df        = original_df,
  label_mapping      = label_mapping_all_flipped,
  max_depth          = 2L,
  output_objects     = c("qini_plot", "diff_gain_summaries")
)

qini_plots <- map(qini_results_valid, ~ .x$qini_plot)

# grab pretty outcome names
qini_names <- margot_get_labels(reliable_ids, label_mapping_all_flipped)

cli_h1("Qini curves generated ✓")
#all good up to here

# 4  POLICY TREES (max depth = 2) ----------------------------------------

policy_results_2L <- margot_policy(
  models_binary_flipped_all,
  save_plots         = FALSE,
  output_dir         = here::here(push_mods),
  decision_tree_args = decision_tree_defaults,
 # policy_tree_args   = policy_tree_defaults,
  model_names        = reliable_ids,      # only those passing Qini
  max_depth          = 2L,
  original_df        = original_df,
  label_mapping      = label_mapping_all_flipped,
  output_objects     = c("combined_plot")
```

```
)

policy_plots <- map(policy_results_2L, ~ .x$combined_plot)


# 5  PLAIN-LANGUAGE INTERPRETATION OF TREES ---------------------------------

policy_text <- margot_interpret_policy_batch(
  models           = models_binary_flipped_all,
  original_df      = original_df,
  model_names      = reliable_ids,
  label_mapping    = label_mapping_all_flipped,
  max_depth        = 2L
)

cat(policy_text, "\n")

cli::cli_h1("Finished: depth-2 policy trees analysed ✓")

# ──────────────────────────────────────────── EOF
──────────────────────────────────────────────────────────────────────────────


# count of valid outcomes
length(qini_names)

# names of valid models

glued_policy_names_1 <- qini_names[[1]]
glued_policy_names_2 <- qini_names[[2]]
# glued_policy_names_3 <- qini_names[[3]]
# glued_policy_names_4 <- qini_names[[4]]
# glued_policy_names_5 <- qini_names[[5]]
# glued_policy_names_6 <- qini_names[[6]]
# glued_policy_names_7 <- qini_names[[7]]


# view qini plots ----------------------------------------------------------

library(patchwork)
# combine first column of plots (4,6,7,8) and second column (9,11,12)
# these showed reliable qini results

combined_qini <- (
  qini_plots[[1]]
) | (
  # remove this block if you don't have plots 9,11,12
  qini_plots[[2]]
) +
  # collect all legends into one shared guide
  plot_layout(guides = "collect") +
```

```
  # add title (and optionally subtitle)
  plot_annotation(
    title    = "Combined Qini Plots",
    subtitle = "Panels arranged with shared legend"
  ) &
  # apply theme modifications to all subplots
  theme(
    legend.position   = "bottom",          # place legend below
    plot.title        = element_text(hjust = 0.5),  # centre title
    plot.subtitle     = element_text(hjust = 0.5)   # centre subtitle
  )

# draw it
print(combined_qini)

# PLANNED COMPARISONS ----------------------------------------------------

# +--------------------------+
# |    MODIFY THIS SECTION   |
# +--------------------------+



# compact example ----------------------------------------------------------

# ------------------------------------------------------------------------------
# purpose: explores effect heterogeneity across researcher-defined strata (appendix only)
# ------------------------------------------------------------------------------
# workflow philosophy ------------------------------------------------------------
# • descriptive, not prescriptive. we *report* how τ̂ varies over groups – we do
#   NOT optimise a policy rule.
# • strata = ±1 sd splits by default; change to suit theory.
# • code chunks are short and labelled so students can run/debug in order.
# ------------------------------------------------------------------------------

library(margot)      # ≥ 1.0.52


library(tidyverse)   # pipes + helpers
library(knitr)       # kable tables
library(patchwork)   # plot stacks

# check package version early ----------------------------------------------------
stopifnot(utils::packageVersion("margot") >= "1.0.52")


#problem from here

# ------------------------- 1. define strata ---------------------------------
# 0. back-transform helper ------------------------------------------
```

```
# margot stores income as z-scored log dollars.  to write interpretable
# subtitles we convert ±1 sd back to the raw scale.  the helper simply
# inverts the: log → z transformation.
log_mean_inc <- mean(original_df$t0_log_household_inc, na.rm = TRUE)

log_sd_inc   <- sd  (original_df$t0_log_household_inc, na.rm = TRUE)

margot_back_transform_log_z(
  log_mean = log_mean_inc,
  log_sd   = log_sd_inc,
  z_scores = c(-1, 0, 1),
  label    = "data_scale"   # prints nz$ values ≈ 41k,…
)

# 1. define strata via logical vectors ------------------------------
# we treat ±1sd as the default cut for “low / mid / high”.  students
# can change the thresholds or supply any logical `subset_condition`.
complex_condition_political <- between(X[, "t0_political_conservative_z"], -1, 1)
complex_condition_wealth    <- between(X[, "t0_log_household_inc_z"], -1, 1)
complex_condition_age       <- between(X[, "t0_age_z"], -1, 1)

# sanity-check age bounds on the raw scale
mean(original_df$t0_age) + c(-1, 1) * sd(original_df$t0_age)

# age ---------------------------------------------------------------------
subsets_standard_age <- list(
  Younger = list(
    var = "t0_age_z",
    value = -1,
    operator = "<",
    label = "Age < 35"
  ),
  Middle = list(
    var = "t0_age_z",
    # operator = "<",
    subset_condition = complex_condition_age,
    label = "Age 35-62"
  ),
  Older = list(
    var = "t0_age_z",
    value = 1,
    operator = ">",
    label = "Age > 62"
  )
)

# gender (binary) ---------------------------------------------------------
subsets_standard_gender <- list(
  Female = list(
    var = "t0_male_binary",
```

```
    value = 0,
    description = "Females"
  ),
  Male = list(
    var = "t0_male_binary",
    value = 1,
    description = "Males"
  )
)

# ethnicity (binary dummies) ------------------------------------------------
subsets_standard_ethnicity <- list(
  Euro = list(
    var = "t0_eth_cat_euro_binary",
    value = 1,
    label = "NZ Europeans ",
    description = "NZ Europeans"


  ),
  Maori = list(
    var = "t0_eth_cat_maori_binary",
    value = 1,
    label = "Māori",
    description = 'Māori'
  )
)
# religion ------------------------------------------------------------------
subsets_standard_secular_vs_religious <- list(
  Not_Religious = list(var = "t0_religion_bigger_denominations_not_rel_binary", value = 1,
label = "Not Religious"),
  Religious = list(var = "t0_religion_bigger_denominations_not_rel_binary", value = 0, label
= "Religious")
)

# ------------------------- 2. defaults for ATE plots -------------------------
# set up domain names
domain_names <- c("wellbeing")

# set up subtitles
subtitles <- ""

# play around with these values
x_offset_comp <- 1.0
x_lim_lo_comp <- -1.0
x_lim_hi_comp <- 1.0

base_defaults_comparisons <- list(
  type = "RD",
  title = ate_title,
```

```
  e_val_bound_threshold = 1.2,
  label_mapping = label_mapping_all,
  adjust = "bonferroni",
  #<- new
  alpha = 0.05,
  # <- new
  colors = c(
    "positive" = "#E69F00",
    "not reliable" = "grey50",
    "negative" = "#56B4E9"
  ),
  x_offset = x_offset_comp,
  # will be set based on type
  x_lim_lo = x_lim_lo_comp,
  # will be set based on type
  x_lim_hi = x_lim_hi_comp,
  text_size = 8,
  linewidth = 0.75,
  estimate_scale = 1,
  base_size = 18,
  point_size = 2.5,
  title_size = 19,
  subtitle_size = 16,
  legend_text_size = 10,
  legend_title_size = 10,
  include_coefficients = FALSE
)

#issues from here

# ------------------------- 3. batch subgroup analysis -------------------------

planned_subset_results <- margot_planned_subgroups_batch(
  domain_models = list(models_binary),
  X             = X,
  base_defaults = base_defaults_binary,
  subset_types  = list(
    ethnicity           = subsets_standard_ethnicity
  ),
  original_df   = original_df,
  label_mapping = label_mapping_all,
  domain_names  = "wellbeing",
  adjust        = "bonferroni",
  alpha         = 0.05,
  subtitles  = subtitles
)



# ethnicity – 2×2 grid
```

```
ethnicity_plot<- wrap_plots(
  list(
    planned_subset_results$wellbeing$ethnicity$results$`NZ Europeans `$plot,
    planned_subset_results$wellbeing$ethnicity$results$Māori$plot
  ),
  ncol = 2
) +
  plot_annotation(
    title = "Ethnicity",
    theme = theme(plot.title = element_text(size = 18, face = "bold"))
  )
ethnicity_plot

cli::cli_h1("subgroup analysis complete ✓")


# example: secular vs religious
group_comparison_secular_religious <- margot_compare_groups(
  group1_name = "NZ EURO",
  group2_name = "Maori",
  planned_subset_results$wellbeing$ethnicity$results$`NZ Europeans `$transformed_table,
  # reference
  planned_subset_results$wellbeing$ethnicity$results$Māori$transformed_table,
  # reference
  type           = "RD",
  # risk-difference scale
  decimal_places  = 3
)
print(group_comparison_secular_religious$results |> kbl("markdown", digits = 2))
cat(group_comparison_secular_religious$interpretation)



# End of Script ------------------------------------------------------



# EXTRA MATERIAL -----------------------------------------------------



# FOR APPENDIX IF DESIRED --------------------------------------------
# helper: combine and save ggplot objects ---------------------------------
combine_and_save <- function(plots, prefix) {
  if (length(plots) == 0) {
    message("no ", prefix, " plots to combine")
    return(invisible(NULL))
  }
  cols     <- ifelse(length(plots) > 3, 2, 1)
  combined <- purrr::reduce(plots, `+`) +
```

```
    patchwork::plot_layout(ncol = cols) &
    patchwork::plot_annotation(
      title    = toupper(prefix),
      subtitle = glue::glue("{length(plots)} models"),
      tag_levels = "A"
    )
  print(combined)
  ggsave(
    here::here(push_mods, paste0("combined_", prefix, ".pdf")),
    combined,
    width  = ifelse(cols == 1, 8, 12),
    height = 6 * ceiling(length(plots) / cols)
  )
  combined
}

# step 3: plot rate curves ------------------------------------------------
autoc_plots <-
  margot_plot_rate_batch(
    models      = models_binary_flipped_all,
    save_plots  = FALSE,
    label_mapping = label_mapping_all,
    model_names = model_groups$autoc
  )
autoc_plots[[1]]

combined_autoc <- combine_and_save(autoc_plots, "rate_autoc")
model_groups$exploratory


# ILLLUSTRATIONS OF SETTINGS
# OPTIONS FOR DECISION TREES ----------------------------------------------
# plot options: showcased -------------------------------------------
# default
margot_plot_decision_tree(models_binary, "model_t2_kessler_latent_anxiety_z", )
# tighten branches for easier viewing in single graphs
margot::margot_plot_decision_tree(
  models_binary,
  "model_t2_kessler_latent_anxiety_z",
  span_ratio = .30,
  text_size = 3.8,
  border_size = .1,
  #  title = "none",
  original_df = original_df
)
# colour decision node
margot::margot_plot_decision_tree(
  models_binary,
  "model_t2_kessler_latent_anxiety_z",
  span_ratio = .3,
```

```
  text_size = 4,
  title = "New Title",
  non_leaf_fill =  "pink",
  original_df = original_df
)
# make new title
margot::margot_plot_decision_tree(
  models_binary,
  "model_t2_kessler_latent_anxiety_z",
  span_ratio = .2,
  text_size = 3,
  title = "New Title",
  non_leaf_fill =  "white",
  original_df = original_df
)


# adjust only the alpha
margot::margot_plot_policy_tree(models_binary, "model_t2_kessler_latent_anxiety_z",
point_alpha = .1)
margot::margot_plot_policy_tree(models_binary, "model_t2_kessler_latent_anxiety_z",
point_alpha = .9)
```